PATENT
Attorney Docket No.: 16869N-104900US
Client Ref. No.: NT1432US

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:

KATSUYA TANAKA et al.

Application No.: 10/770,723

Filed: February 2, 2004

For: STORAGE DEVICE AND CONTROLLING METHOD THEREOF

Customer No.: 20350

Examiner: Unassigned

Technology Center/Art Unit: 2655

Confirmation No.: 2283

PETITION TO MAKE SPECIAL FOR NEW APPLICATION UNDER M.P.E.P. § 708.02, VIII & 37 C.F.R. § 1.102(d)

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

This is a petition to make special the above-identified application under MPEP § 708.02, VIII & 37 C.F.R. § 1.102(d). The application has not received any examination by an Examiner.

(a) The Commissioner is authorized to charge the petition fee of $130 under 37 C.F.R. § 1.17(i) and any other fees associated with this paper to Deposit Account 20-1430.

(b) All the claims are believed to be directed to a single invention. If the Office determines that all the claims presented are not obviously directed to a single invention, then Applicants will make an election without traverse as a prerequisite to the grant of special status.

(c) Pre-examination searches were made of U.S. issued patents, including a classification search and a key word search. The classification search was conducted on or around April 1, 2005 covering Class 369 (subclasses 47.15 and 47.36), Class 710 (subclass 316), and Class 711 (subclasses 101, 111, 112, 113, 114, 154, and 170), by a professional search firm, Lacasse & Associates, LLC. The key word search was performed on the USPTO full-text database including published U.S. patent applications. The inventors further provided four references considered most closely related to the subject matter of the present application (see references #7-10 below), which were cited in the Information Disclosure Statement filed with the application on February 2, 2004.

(d) The following references, copies of which are attached herewith, are deemed most closely related to the subject matter encompassed by the claims:

(1) U.S. Patent No. 6,295,587 B1;

(2) U.S. Patent No. 6,587,919 B2;

(3) U.S. Patent No. 6,640,281 B2;

(4) U.S. Patent No. 6,701,411 B2;

(5) U.S. Patent Publication No. 2003/0191891 A1;

(6) U.S. Patent Publication No. 2005/0027919 A1;

(7) Japanese Patent Publication No. JP 2003-85117;

(8) Japanese Patent Publication No. JP 2000-222339;

(9) Japanese Patent Publication No. JP 2003-303055; and

(10) Qlogic Corp., "Full Duplex and Fibre Channel, Network Storage Group Host Products Technology Brief."

(e) Set forth below is a detailed discussion of references which points out with particularity how the claimed subject matter is distinguishable over the references.

A. Claimed Embodiments of the Present Invention

The claimed embodiments relate to controlling data transfer of a storage device.

Independent claim 1 recites a disk device comprising a disk controller comprising a channel adapter, a cache memory, and a disk adapter; and a disk array comprising disk drives, each being equipped with a plurality of I/O ports. The disk adapter and the disk array are connected via a switch. A destination drive I/O port to which a frame is to be forwarded is determined, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives.

Independent claim 4 recites a disk device comprising a disk controller comprising a channel adapter, a cache memory, and a disk adapter; and a disk array comprising disk drives, each being equipped with a plurality of I/O ports. The disk adapter and the disk array are connected via a switch. A path which a frame passes to be transferred between the switch and one of the disk drives is determined, according to the type of a command included in an exchange between the disk adapter and the one of the disk drives.

Independent claim 6 recites a disk device comprising a disk controller comprising a channel adapter, a cache memory, and a disk adapter; and a disk array comprising disk drives, each being equipped with a plurality of I/O ports. The disk adapter and the disk array are connected via a switch. The disk adapter determines destination information within a frame to be transferred from the disk adapter to one of the disk drives, according the type of a command included in an exchange between the disk adapter and the one of the disk drives. The switch selects one of port to port connection paths between a port to which the disk adapter is connected and ports to which the disk drives constituting the disk array are connected to switch each frame inputted to the switch, according to the destination information within the frame.

Independent claim 7 recites a disk device comprising a disk controller comprising a channel adapter, a cache memory, and a disk adapter; and a disk array comprising disk drives, each being equipped with a plurality of I/O ports. The disk adapter and the disk array are connected via a switch. A destination drive port to which a frame is to be forwarded is determined, depending on whether the type of a command included in an

exchange that is transferred between the disk adapter and one of the disk drives is a data read command or a data write command. The exchange for reading data and the exchange for writing data are executed in parallel.

Independent claim 8 recites a disk device comprising a disk controller comprising a channel adapter, a cache memory, and a disk adapter; and a disk array comprising disk drives, each being equipped with a plurality of I/O ports. The disk adapter and the disk array are connected via a switch. A path which a frame passes between the switch and one of the disk drives is determined, depending on whether the type of a command included in an exchange between the disk adapter and the one of the disk drives is a data read command or a data write command.

Independent claim 9 recites a disk device comprising a disk controller comprising a channel adapter, a cache memory, and a disk adapter; a plurality of disk drives, each being equipped with a plurality of I/O ports; and a switch connecting the disk controller and the plurality of disk drives. A destination drive port to which a frame is to be forwarded is determined, depending on whether the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives is a data read command or a data write command. The exchange for reading data and the exchange for writing data are executed in parallel.

One of the benefits that may be derived is that a disk device having a back-end network that enables full duplex data transfer by simple control techniques an be realized, and that increased disk device throughput is achieved.

B. Discussion of the References

None of the following references disclose determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9. For instance, claim 1 recites a destination drive I/O port to which a frame is to be forwarded is determined, according to the type of a command included in an exchange that is transferred between the disk adapter

and one of the disk drives. Claim 4 recites a path which a frame passes to be transferred between the switch and one of the disk drives is determined, according to the type of a command included in an exchange between the disk adapter and the one of the disk drives. Claim 6 recites that the disk adapter determines destination information within a frame to be transferred from the disk adapter to one of the disk drives, according the type of a command included in an exchange between the disk adapter and the one of the disk drives. Claim 7 recites a destination drive port to which a frame is to be forwarded is determined, depending on whether the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives is a data read command or a data write command. Claim 8 recites a path which a frame passes between the switch and one of the disk drives is determined, depending on whether the type of a command included in an exchange between the disk adapter and the one of the disk drives is a data read command or a data write command. Claim 9 recites a destination drive port to which a frame is to be forwarded is determined, depending on whether the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives is a data read command or a data write command.

1. U.S. Patent No. 6,295,587 B1

The patent to Martin (6,295,587 B1), assigned to EMC Corp., provides for a Method and Apparatus for Multiple Disk Drive Access in a Multi-Processor/Multi-Disk Drive System. Disclosed is disk device 14, drive adapters 18, and adapter ports (AP) 20 and drive ports (DP) 24 of switch 22. AP 20 of switch 22 is connected to drive adapters 18 and DP 24 of switch 22 is connected to the input/output ports of disk devices 14. Disk identifier (ID) 48 and switch state (SS) 50 may control both the adapter 18 and switch 22 to establish appropriate connection between processor 12 and disk device 14. See column 3, line 48 to column 5, line 39; and Figure 1.

This reference relates to the use of a switch with a binding mapper and an address mapper. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an

exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

2. U.S. Patent No. 6,587,919 B2

The patent to Yanai et al. (6,587,919 B2), assigned to EMC Corp., provides for a System and Method for Disk Mapping and Data Retrieval. Disclosed is disk storage system 10 with means for receiving write commands and data such as channel adapter boards 12a-d, which receive disk read/write commands and data over communication channels 1-8. Disk adapter boards 20 read and write data to one or more disk drive units 22. Channel adapter boards 12a-d are connected to cache memory storage unit 16. When channel adapter boards 12a-d receive write commands, there is stored in memory an indication to disk adapters 20 that data record stored in cache must be written to the disk drives. See column 5, line 52 to column 6, line 18.

This reference relates to the use of a record locator data structure having variable-length data records for disk mapping and data retrieval. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

3. U.S. Patent No. 6,640,281 B2

The patent to Obara et al. (6,640,281 B2), assigned to Hitachi, Ltd., provides for Storage Subsystem with Management Site Changing Function. Disclosed is disk controller 8, host interfaces 2, channel paths 9, cache memory 4, disk interfaces 5, and plurality of disk drives 6. Cache memory temporarily stores data written in disk controller and data read from disk drive 8 and output to the host. Data exchange between host computer 1 and disk controller is performed via channel path 9, where data is called a frame. Frame received at host interface 2 is identified for target volume and type of operation from the volume number field 24 and command/data field 25, where the input/output command is sent

to head disk assemblies (HDA) of disk drive. For read command, HDA returns the read data to the disk interface. See column 4, line 30 to column 6, line 20; and Figures 1 and 2.

This reference relates to a technique for transferring management of desired disk drives and volumes to be managed by an overload disk controller under access concentration to an optional disk controller not in an overload state while an application on the host and an ordinary process of the disk controllers are maintained to continue. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

4. U.S. Patent No. 6,701,411 B2

The patent to Matsunami et al. (6,701,411 B2), assigned to Hitachi, Ltd., provides for a Switch and Storage System for Sending an Access Request from a Host to a Storage Subsystem. Disclosed is host adapter 101, diskarray interface (I/F) controller 1011 to control the diskarray switches 20, diskarray I/F 21, and host bus 1012 to perform communications and data transfer between cache memory 102 and diskarray I/F controller 1011. Lower adapter 103 executes control of disk I/F controller 1031 to control disk 104 and disk I/F. Diskarray switch 20 contains managing processor (MP) to perform functions such as management and control of diskarray switch and crossbar switch 201. During a read operation, switching controller (SC) 2022 reads the frame held in frame buffer (FB) 2021 and analyzes the frame header 401. Information such as ID for the frame transfer destination are included in the frame header. See column 4, line 55 to column 5, line 17; column 9, lines 27-45; and Figures 1 and 2.

This reference relates to the use of a switch connected between a first interface node and a plurality of second interface nodes to perform frame transfer between the first interface node and the second interface nodes based on node address information added to the frame. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of

the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

5. U.S. Patent Publication No. 2003/0191891 A1

The patent application publication to Tanaka et al. (2003/0191891 A1), assigned to Hitachi, Ltd., provides for a Disk Storage System having Disk Arrays Connected with Disk Adaptors through Switches. Disclosed is disk storage system comprising disk adapter DKA performing control required when data is transmitted and received between disk controller DKC and disk array DA, wherein DKA is connected to DA through channels D01-04. C1-4 are channels allowing communication between channel adapters CHA and CPU. Cache memory CM functions to temporarily store data inputted/outputted by disk storage system. Switch SW1 includes input/output ports P1-P5. During writing of data in DA, circuit configuration allows a frame in a block to be inputted from port P1 and outputted from ports P2-5. See paragraphs [0071], [0073], [0074], [0075], [0092], [0097], and [0098].

This reference discloses the use of switches between a disk adapter and a disk array to change over connection between ports to which the disk adapter is connected and ports to which disk drives constituting the disk array are connected in accordance with destination information in a frame for each of the inputted frames. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

6. U.S. Patent Publication No. 2005/0027919 A1

The patent application publication to Aruga (2005/0027919 A1) provides for a Disk Subsystem. Disclosed is N disk array controllers 1-1 to 1-N, each with M disk drive interface controllers 2-1 to 2-M, where each of the M controllers of fibre channel fabric switch 3-1 to 3-M are respectively connected to the disk drive interface controllers 2-1 to 2-

M for controlling disk drive units through fibre channel interface 5. Data transferred between host computer and disk array are temporarily stored in cache memory. Switch controller 17 sets switch 18 based on the ID number received from disk drive interface controllers 2-1 to 2-M. See paragraphs [0023], [0025], and [0029]; and Figure 1.

This reference relates to the use of a protocol controller disposed between switches in a fiber channel fabric switch circuit and disk drive units for converting a protocol to enable one-to-one connectivity established between controllers and disk drive units. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

7. Japanese Patent Publication No. JP 2003-85117

This reference provides a storage controlling device capable of efficiently using a full duplex type communication passage. The storage controlling device connects to an upward processing device through the full duplex type communication passage and stores and manages data received through the communication passage into a data storing means. The storage controlling device has a plurality of channel processors for inputting or outputting data from a data storing means in response to a command included in the data (frame) transmitted from the upward processing device, and assigns the channel processors for inputting or outputting the data related to the data (frame) in response to the command included in the data (frame).

As described in the present application at page 3, line 18 to page 4, line 2, the reference discloses that channel processors for inputting data to and outputting data from the disk device are controlled in accordance with a command from the host device and the quantity of data to be transferred so that full duplex operation is performed between the host device and the storage controlling device. However, dynamic control is required when data is transferred and its problem is complexity of the control method. Also, the reference does not deal with the full duplex data transfer in the back-end of a disk device. See page 5, lines 17-21.

This reference does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

8. Japanese Patent Publication No. JP 2000-222339

This reference discloses a way to connect plural disk drives with a disk drive interface circuit without sacrificing transmitting performance by using a fiber channel fabric topology for reducing the number of connection lines by using a fiber channel interface as a serial interface, and for realizing switch connection. A fiber channel switch control circuit 3 is provided between a disk drive 4 and a disk drive interface control circuit 2, and protocol control part 16 is provided between a switch 18 in the fabric switch circuit 3 and the disk drive.

As described in the present application at page 4, lines 3-8, the reference discloses a disk array system where a disk array controller and disk drives are connected via a switch. The reference, however, does not deal with application of the technique to the back-end of a disk drive equipped with a plurality of I/O ports and the full duplex data transfer in the back-end. See page 5, lines 22-25.

This reference does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

9. Japanese Patent Publication No. JP 2003-303055

This reference discloses a disk storage system having high throughput between a disk adapter of a disk controller and a disk array. The disk adapter of the disk controller is connected to the disk array through switches. Data on a channel between the switch and a RAID group is multiplexed in the switch to be transferred onto a channel between the switch

and the disk adapter, and data on the channel between the switch and the disk adapter is demultiplexed in the switch to be transferred onto the channel between the switch and the RAID group. A data transfer rate on the channel between the disk adapter and the switch is made higher than that on the channel.

This reference relates to the use of a switch with multiplexing and demultiplexing to achieve higher throughput between a disk adapter of a disk controller and a disk array. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

10. Qlogic Corp., "Full Duplex and Fibre Channel, Network Storage Group Host Products Technology Brief."

This reference discusses the meaning of full-duplex data transmission with fibre channel.

As discussed in the present application at page 3, lines 5-17, the reference discloses a plurality of FC-ALs in which disk drives are connected and a server are connected via a switch and parallel data transfers are carried out between the server and the plurality of FC-ALs. It does not take a disk drive having a plurality of I/O ports into consideration and it is difficult to apply the technique to a disk device comprising disk drives each having a plurality of I/O ports in the back-end. See page 5, lines 10-16.

This reference does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

(f) In view of this petition, the Examiner is respectfully requested to issue a first Office Action at an early date.

Respectfully submitted,

Chun-Pok Leung
Reg. No. 41,405

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 650-326-2400
Fax: 415-576-0300
Attachments
RL:rl
60475672 v1

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-085117
(43)Date of publication of application : 20.03.2003

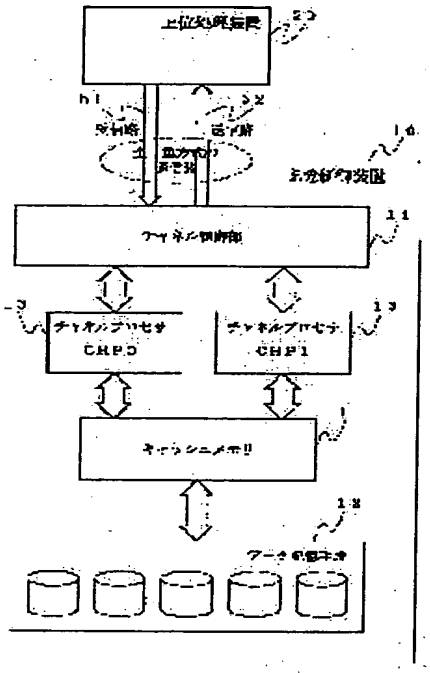
(51)Int.CI. G06F 13/10
G06F 3/06
G06F 13/12

(21)Application number : 2001-273932
(22)Date of filing : 10.09.2001

(71)Applicant : HITACHI LTD
(72)Inventor : MAEDA MASAMI
AZUMI YOSHIHIRO
SAKAKI TOSHINORI
TSUKADA MASARU

(54) STORAGE CONTROLLING DEVICE, AND ITS OPERATING METHOD

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a storage controlling device capable of efficiently using a full duplex type communication passage.
SOLUTION: This storage controlling device connects to an upward processing device through the full duplex type communication passage and stores and manages data received through the communication passage into a data storing means. The storage controlling device has a plurality of channel processors for inputting or outputting data from a data storing means in response to a command included in the data (frame) transmitted from the upward processing device, and assigns the channel processors for inputting or outputting the data related to the data (frame) in response to the command included in the data (frame).



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2003－85117
（P2003－85117A）
(43)公開日　平成15年３月20日(2003.3.20)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　13/10	３４０	Ｇ０６Ｆ　13/10	３４０Ａ	５Ｂ０１４
3/06	３０１	3/06	３０１Ａ	５Ｂ０６５
13/12	３４０	13/12	３４０Ｃ	

審査請求　未請求　請求項の数14　ＯＬ　（全 14 頁）

(21)出願番号　特願2001－273932(P2001－273932)

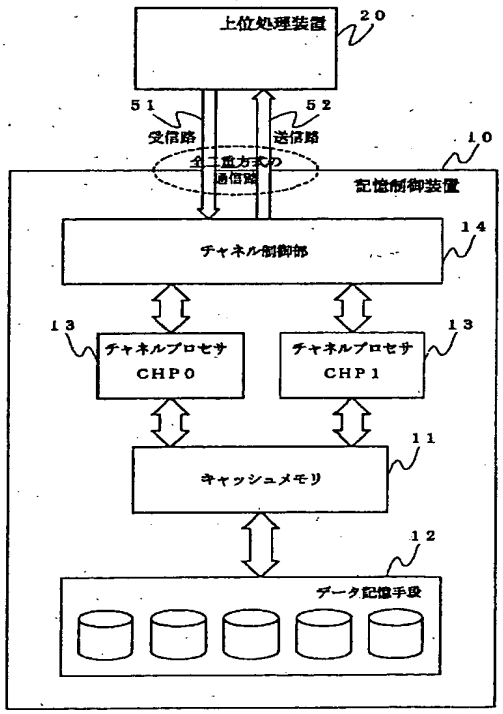
(22)出願日　平成13年９月10日(2001.9.10)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　前田　昌美
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(72)発明者　安積　義弘
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(74)代理人　100071283
弁理士　一色　健輔　（外５名）

最終頁に続く

(54)【発明の名称】　記憶制御装置およびその運用方法

(57)【要約】
【課題】　全二重方式の通信路を効率よく利用することができる記憶制御装置を提供する。
【解決手段】　全二重方式の通信路により上位処理装置と接続し、通信路を通じて受信したデータをデータ記憶手段に記憶管理する記憶制御装置において、上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応してデータ記憶手段に対するデータ入出力処理を行う複数のチャネルプロセサを備え、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関する前記データ入出力処理を実行するチャネルプロセサを割り当てるようにする。



【特許請求の範囲】

【請求項１】　全二重方式の通信路により上位処理装置と接続し、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理する記憶制御装置であって、
前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャネルプロセサを備え、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関する前記データ入出力処理を実行するチャネルプロセサを割り当てる手段を備えることを特徴とする記憶制御装置。

【請求項２】　請求項１に記載の記憶制御装置であって、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関するデータ入出力処理を実行する前記チャネルプロセサを割り当てる前記手段が、前記データ（フレーム）に含まれる前記コマンドが前記データ記憶手段にデータの書き込みコマンドであるかデータの読み出しコマンドであるかに応じて前記データ（フレーム）についての処理を行うチャネルプロセサを割り当てる手段であることを特徴とする記憶制御装置。

【請求項３】　請求項１に記載の記憶制御装置であって、前記通信路の送信路と受信路を流れるデータ量に応じて前記データ（フレーム）についての処理を行うチャネルプロセサを割り当てる手段を備えることを特徴とする記憶制御装置。

【請求項４】　請求項１に記載の記憶制御装置であって、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータフレームに関するデータ入出力処理を実行する前記チャネルプロセサを割り当てる前記手段を実行するかどうかを、前記データ記憶手段に対するデータの書き込みコマンドとデータの読み出しコマンドにより単位時間内に処理されたデータ量に応じて制御する手段を備えることを特徴とする記憶制御装置。

【請求項５】　請求項１に記載の記憶制御装置であって、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関するデータ入出力処理を実行する前記チャネルプロセサを割り当てる前記手段を実行するかどうかを、前記各チャネルプロセサの処理待ちキューにキューイングされているデータ書き込みコマンドの数とデータ読み出しコマンドの数に応じて制御する手段を備えることを特徴とする記憶制御装置。

【請求項６】　請求項１に記載の記憶制御装置であって、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関するデータ入出力処理を実行する前記チャネルプロセサを割り当てる前記手段を実行するかどうかを、前記各チャネルプロセサが前記データ入出力処理において単位時間内に処理したデータ量に応じて制御する手段を備えることを特徴とする記憶制御装置。

【請求項７】　請求項１に記載の記憶制御装置であって、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関するデータ入出力処理を実行する前記チャネルプロセサを割り当てる前記手段を実行するかどうかを、当該記憶制御装置における、前記データ記憶手段に対するデータ書き込みコマンドの処理についてのスループットと、前記データ記憶手段に対するデータ読み出しコマンドの処理についてのスループットに応じて制御する手段を備えることを特徴とする記憶制御装置。

【請求項８】　請求項４または６に記載の記憶制御装置であって、前記単位時間を当該記憶制御装置に接続された外部装置から指定させる手段を備えることを特徴とする請求項４または６に記載の記憶制御装置。

【請求項９】　全二重方式の通信路により上位処理装置に接続され、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理し、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャネルプロセサを備えて構成される記憶制御装置の運用方法であって、
前記データ（フレーム）に含まれる前記コマンドが前記データ記憶手段にデータの書き込みコマンドであるかデータの読み出しコマンドであるかに応じて前記データ（フレーム）についての処理を行うチャネルプロセサを割り当てるようにしたことを特徴とする記憶制御装置の運用方法。

【請求項１０】　全二重方式の通信路により上位処理装置に接続され、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理し、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャネルプロセサを備えて構成される記憶制御装置の運用方法であって、
前記通信路の送信路と受信路を流れるデータ量に応じて前記データ（フレーム）についての処理を行うチャネルプロセサを割り当てるようにしたことを特徴とする記憶制御装置の運用方法。

【請求項１１】　全二重方式の通信路により上位処理装置に接続され、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理し、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャネルプロセサを備えて構成される記憶制御装置の運用方法であって、
前記データ（フレーム）に含まれるコマンドの種類と、前記データ記憶手段に対するデータの書き込みコマンドとデータの読み出しコマンドにより単位時間内に処理されたデータ量とに応じてそのデータ（フレーム）に関す

る前記データ入出力処理を実行するチャネルプロセサを割り当てるようにしたことを特徴とする記憶制御装置の運用方法。

【請求項１２】　全二重方式の通信路により上位処理装置に接続され、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理し、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャネルプロセサを備えて構成される記憶制御装置の運用方法であって、
前記データ（フレーム）に含まれるコマンドの種類と、前記各チャネルプロセサの処理待ちキューにキューイングされているデータ書き込みコマンドの数とデータ読み出しコマンドの数とに応じてそのデータ（フレーム）に関する前記データ入出力処理を実行するチャネルプロセサを割り当てるようにしたことを特徴とする記憶制御装置の運用方法。

【請求項１３】　全二重方式の通信路により上位処理装置に接続され、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理し、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャネルプロセサを備えて構成される記憶制御装置の運用方法であって、
前記データ（フレーム）に含まれるコマンドの種類と、前記各チャネルプロセサが前記データ入出力処理において単位時間内に処理したデータ量とに応じてそのデータ（フレーム）に関する前記データ入出力処理を実行するチャネルプロセサを割り当てるようにしたことを特徴とする記憶制御装置の運用方法。

【請求項１４】　全二重方式の通信路により上位処理装置に接続され、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理し、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャネルプロセサを備えて構成される記憶制御装置の運用方法であって、
前記データ（フレーム）に含まれるコマンドの種類と、当該記憶制御装置における、前記データ記憶手段に対するデータ書き込みコマンドの処理についてのスループットと、前記データ記憶手段に対するデータ読み出しコマンドの処理についてのスループットとに応じてそのデータ（フレーム）に関する前記データ入出力処理を実行するチャネルプロセサを割り当てるようにしたことを特徴とする記憶制御装置の運用方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、全二重方式の通信路により上位処理装置と接続し、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理する記憶制御装置に関し、とくに、前記通信路を効率よく利用するための技術に関する。

【０００２】

【従来の技術】メインフレームなどの上位処理装置と、ディスクアレイ装置などの記憶制御装置との間を結ぶ通信プロトコルとして、ファイバチャネルプロトコル（以下、「ファイバチャネル」と称する）が知られている（詳細はＡＮＳＩ(American National Standard for Information Technology)のＦＣ－ＰＨ（Fibre Channel Physical and Signaling Protocol）を参照）。

【０００３】ファイバチャネルは、２ポート間の物理接続が基本であり、一対のファイバチャネルポートは、物理的には送受信２本（全二重方式）の通信路で相互に接続されている。この通信路を通じて行われるファイバチャネルにおける記憶制御装置と上位処理装置との間の通信は、フレームと呼ばれるデータ単位を基本として行われる。また、複数のフレームの束はシーケンスと呼び、さらに、シーケンスの束はエクスチェンジと呼ぶ。例えば、上位処理装置から記憶制御装置に対してデータ読み出し命令（Read命令）に対応する一連の処理はエクスチェンジを単位として行われる。また、ファイバチャネルにおける上位処理装置と記憶制御装置との間の通信では、インターロックを取らずにコマンドやフレームを送受信することが可能である。

【０００４】図１に、ファイバチャネルの通信路により接続された記憶制御装置１０と、これを利用する上位処理装置２０とを備えて構成されるデータ処理システムの一例を示す。記憶制御装置１０は、例えば、ディスクアレイ装置などであり、キャッシュメモリ１１、ディスクユニットなどのデータ記憶手段１２、ＣＰＵやメモリなどを主体として構成される。そして、記憶制御装置１０は、上位処理装置２０との間のデータ通信や処理対象となるデータのチャネルプロセサ１３への割り当ておよび各種データやテーブルの管理や、チャネルプロセサ１３への各種命令のキューの管理などを行うチャネル制御部１４、チャネル制御部１４から受信したフレームを切り分けて各フレームに対応する処理やキャッシュメモリ１１を通じたデータ記憶手段１２へのデータ転送制御を実行するチャネルプロセサ１３などを備えている。一方、上位処理装置２０は、例えば、メインフレームやオフコン、パーソナルコンピュータなどである。

【０００５】記憶制御装置１０と上位処理装置２０との間の通信において、記憶制御装置１０は、受信したフレームの順に各フレームを処理するチャネルプロセサ１３を割り当てる。また、この際にチャネルプロセサ１３が使用中の場合には、使用されていないチャネルプロセサ１３がそのフレームの処理用に割り当てられることになる。

【０００６】

【発明が解決しようとする課題】ところで、上位処理装

置２０から記憶制御装置１０に送られてくるフレームに含まれるコマンドは、主としてデータ記憶手段１２へのデータの書き込みを指示するWRITEコマンドと読み出しを指示するREADコマンドに大別され、これらコマンドに対応する処理に際しては、WRITEコマンドもしくはREADコマンドの一方の種類のコマンドのみを含んだフレームが連続して上位処理装置２０から送られてきた場合や、書き込みや読み出しの対象となるデータ量が大きいなどの理由により、複数のチャネルプロセサ１３の双方が同時にWRITEコマンドもしくはREADコマンドのどちらか一方の処理のみを行っている期間が生じることがあり、この期間中は全二重通信が有効に機能しないことになる。

【０００７】すなわち、例えば、チャネルプロセサ１３が２つのみで構成されている場合には、前述した全二重方式の通信路のうち上位処理装置２０から記憶制御装置１０方向の通信路５１（以下、「受信路」と称する）のみの通信負荷が高くなっているにもかかわらず、記憶制御装置１０から上位処理装置２０方向の通信路５２（以下、「送信路」と称する）は殆ど利用されていないため、この期間中は全二重方式の通信路５０が有効に機能しないことになる。

【０００８】本発明はこのような事情に鑑みてなされたもので、全二重方式の通信路を効率よく利用することができる記憶制御装置を提供することを目的とする。

【０００９】

【課題を解決するための手段】この目的を達成するための、本発明の主たる発明は、全二重方式の通信路により上位処理装置と接続し、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理する記憶制御装置であって、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャネルプロセサを備え、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関する前記データ入出力処理を実行するチャネルプロセサを割り当てる手段を備えることとする。

【００１０】具体的には、例えば、ファイバチャネルにより接続された上位処理装置と記憶制御装置との通信の場合であれば、前述のインターロックが不要であるというファイバチャネルの特徴を利用して、フレームに含まれているコマンドが前記データ記憶手段に対する書き込み命令であるのか、読み出し命令であるのかに応じてそのフレームが所属するエクスチェンジを処理するチャネルプロセサをエクスチェンジごとに割り当てる。これにより、全てのチャネルプロセサが同時に書き込みもしくは読み出しの一方の処理のみを実行する期間が減り、これにより全二重方式の通信路の送信路と受信路の負荷バランスの均一化が図られ、通信路を効率よく利用することが可能になる。

【００１１】【発明の実施の形態】以下、本発明の一実施例によるデータ処理システムについて説明する。データ処理システムの構成は図１と同様であるので詳細な説明は省略し、以下では本発明の特徴的な部分を中心に説明する。また、説明の便宜のため、この実施例で説明する記憶制御装置１０は、２つのチャネルプロセサ（以下、「ＣＨＰ」と称する）１３のみを備えるものとし、これらチャネルプロセサをＣＨＰ０、ＣＨＰ１と称することとする。また、以下の説明では「WRITE」を「WR」と、「READ」を「RD」と、それぞれ略記する。

【００１２】記憶制御装置１０のチャネル制御部１４のメモリ上には、上位処理装置２０との間のデータ通信や、データ記憶手段１２に対するデータの読み書き処理などの、上位処理装置２０から送られてくるコマンドに応じて行われる処理に利用される各種テーブルが記憶管理されている。図２はこのうちのエクスチェンジ管理テーブル２００である。このテーブル２００には、上位処理装置２０との間の通信において生成されたエクスチェンジが管理されている。

【００１３】このテーブルのテーブル有効フラグ２０２には、該当のエクスチェンジが現在処理中であるかどうかを示すビットがセットされ、処理中のエクスチェンジについては「１」が、処理中でないエクスチェンジについては「０」がセットされる。デバイス番号２０３には、そのエクスチェンジの処理対象となるデータ記憶手段の記憶媒体（例えば、ディスクユニット）の識別番号がセットされる。ＣＨＰ番号２０５には当該エクスチェンジの処理を行うＣＨＰの番号がセットされる。また、通信領域ポインタ２０４には受信路５１を通じて受信したフレームが格納されているメモリ上の物理アドレスが、ＯＸ－ＩＤ（Originator Exchange_ID）２０７には上位処理装置２０側で割り当てられたエクスチェンジ番号がセットされ、Ｓ－ＩＤ（Source_ID）２０８には送信元のファイバチャネルのポートアドレスが、Ｄ－ＩＤ（Destination_ID）２０９には該当エクスチェンジの送信先となるファイバチャネルのポートアドレスが、それぞれセットされる。

【００１４】なお、以上の項目のうちテーブル有効フラグ以外の項目には、エクスチェンジが処理中である場合、すなわち、テーブル有効フラグが「１」の場合に値がセットされる。また、ＯＸ－ＩＤ２０７、Ｓ－ＩＤ２０８、Ｄ－ＩＤ２０９の値は、各エクスチェンジに一意に対応づけられ、例えば、上位処理装置２０から連続してフレームが送られてきた場合に各フレームがどのエクスチェンジに所属しているか、フレームが先頭フレームであるか（エクスチェンジ管理テーブルにフレームが登録されていない場合）といったことは、各フレームにセットされているＯＸ－ＩＤ２０７、Ｓ－ＩＤ２０８、Ｄ－ＩＤ２０９に一致するエクスチェンジをエクスチェンジ管理テーブル２００から調べることで把握できる。

【００１５】一方、図３は記憶制御装置１０のメモリ上に記憶管理されているＣＨＰ割当処理管理テーブル３００である。このテーブルのＣＨＰ割当処理実行フラグ３０１は、ＣＨＰの割当処理の実行制御に用いられる。このフラグの用途については後述する。

【００１６】受信WRコマンドデータ量カウンタ３０２および受信RDコマンドデータ量カウンタ３０３には、チャネル制御部１４が上位処理装置２０からコマンドを受信した場合に、WRコマンドもしくは、RDコマンドにより処理されたデータ量が加算される。また、送信RDコマンドデータ量カウンタ３０４には、上位処理装置２０から記憶制御装置１０に送信されたRDコマンドに応答して記憶制御装置１０から上位処理装置２０に送信されるデータ量がセットされ、ＣＨＰ１３から処理完了通知があり、あるエクスチェンジがエクスチェンジ管理テーブル２００において無効化された際に値が加算される。

【００１７】ＣＨＰ０データ量カウンタ３０５、およびＣＨＰ１データ量カウンタ３０６には、各ＣＨＰ０，１が処理したデータのデータ量が加算される。RD／WRデータ比率閾値等の各種閾値３０７～３１１は、後述するＣＨＰの割り当て処理の実行制御のために利用されるものである。モニタリング実行間隔タイマ３１２、モニタリング実行間隔Ｉ／Ｏ数３１３は、記憶制御装置１０が実行する各種ポーリング処理に際して参照されるパラメータである。また、モニタリング開始時刻３１４とモニタリング終了時刻３１５は、記憶制御装置１０により処理されたデータのスループットを算出する場合に用いられる。Ｉ／Ｏ数カウンタ３１５は、Ｉ／Ｏ数でモニタリングするのに必要な情報であり、これには上位処理装置２０から送られてきたＩ／Ｏ要求数が加算される。ワークエリア３１６は、各種計算やデータの一時的な保存などに利用される。

【００１８】図４は、チャネル制御部１４のメモリ上に記憶管理されているRD／WRコマンドキュー管理テーブル４０１，４１０である。このテーブル４０１，４１０には、記憶制御装置２０の処理対象となるRDコマンドもしくはWRコマンドのキューイング状態が、ＦＩＦＯ（First In First Out）方式で管理され、キューイング数を示すカウンタ４０２、先頭キューの格納アドレスを示すＩＮポインタ４０３、末尾キューの格納アドレスを示すＯＵＴポインタ４０４、キュー全体のデータサイズを管理するデータ転送量４０５、コマンドごとに対応するキュー管理情報４０６などが管理される。キュー管理情報４０６には、記憶制御装置２０が割り当てたエクスチェンジ番号４０７とデータ転送量４０８、このエクスチェンジを実行するＣＨＰの番号４０９などが記述されている。

【００１９】つぎに、図５に示すフローチャートとともに、記憶制御装置１０と上位処理装置２０との間の全二重方式の通信において記憶制御装置１０が行う、ＣＨＰの割当処理について詳述する。

【００２０】全二重方式の通信路のうち受信路５１を通じて上位処理装置２０からフレームを受信した場合、記憶制御装置１０は、まず、そのフレームが新たなエクスチェンジの起動先頭フレームであるかどうかを調べる（５０１，５０２）。ここでフレームが起動先頭フレームであった場合には、未使用のエクスチェンジ番号を利用してそのフレームに対応するフィールドをエクスチェンジ管理テーブルに新たに登録する（５０３）。なお、フレームが起動先頭フレームでない場合の処理については後述する。

【００２１】つぎに記憶制御装置１０は、前記フレームのフレーム制御フィールド（F_ControlField）を参照し、当該フレームの後続フレームの存在有無を調べる（５０７）。その結果、当該フレームが後続フレームを有する場合には、さらにこのフレームがRD／WRいずれかのコマンドを含むフレームであるかどうかを調べる（５０８）。なお、この調査は、例えば、記憶制御装置１０内にあらかじめ登録しておいたコマンド一覧と、フレームのコマンド記述欄の内容を比較することで行う。

【００２２】この調査の結果、このフレームがコマンドも含んでいない場合（もしくは、コマンドを含んでいるかどうかを判断できない場合）には、記憶制御装置１０は、当該フレームを通常のＣＨＰ割当方式、すなわち、受信したフレームの順に各フレームを処理するチャネルプロセサ１３を割り当て、また、あるフレームの割り当てに際してチャネルプロセサ１３が使用中である場合には、使用されていないチャネルプロセサ１３にそのフレームの処理させるというＣＨＰ割当方式により当該フレームの処理を行う（５１０）。

【００２３】一方、前記フレームがRD／WRいずれかのコマンドを含むフレームである場合には、つぎのように処理が行われる。まず、フレームにWRコマンドが含まれていた場合、記憶制御装置１０は、WRデータキュー管理テーブル４０１に当該フレームのエクスチェンジ番号とそのWRコマンドにより処理されるデータ量を登録する（５１１）。一方、当該フレームにRDコマンドが含まれていた場合には、記憶制御装置はRDデータキュー管理テーブルに当該フレームのエクスチェンジ番号と、そのREADコマンドにより処理されるデータ量を登録する（５１２）。なお、これら各キュー管理テーブルへの登録が行った場合には、各テーブルのカウンタ４０２およびＯＵＴポインタ４０４に１を加算する。

【００２４】つぎに記憶制御装置１０は、条件管理テーブルのＣＨＰ割当処理実行フラグ３０１の状態を調査する。ここでＣＨＰ割当処理実行フラグ３０１に「１」がセットされていた場合には、ＣＨＰの割当処理を行うかどうかの判断する処理に進み（５１３，５１４）、当該フレームにWRコマンドが含まれていた場合は、エクスチェンジ管理テーブル２００の当該フレームが所属するエ

クスチェンジのＣＨＰ番号２０５の欄に、そのエクスチェンジにＣＨＰ０を割り当てたことを示す「０」をセットし（５１５）、また、ＣＨＰ割当処理管理テーブル３００のＣＨＰ０データ量カウンタ３０５に当該フレームのデータ量を加算し（５１６）、当該フレームの処理を行うため当該フレームについての処理に必要なデータをＣＨＰ１に転送する（５１９）。なお、フレームにRDコマンドが含まれていた場合も、以上のWRコマンドの場合と同じようにして処理が行われる（５１２，５１４，５１７，５１８）。

【００２５】一方、（５１３，５１４）の処理において、ＣＨＰ割当処理実行フラグ３０１に「０」がセットされていた場合には、記憶制御装置１０はコマンドの種類（WRコマンドであるかRDコマンドであるか）に応じたＣＨＰの割当処理を行わず、その代わりにＣＨＰ割当処理管理テーブル３００のＣＨＰ０データ量カウンタ３０５もしくはＣＨＰ１データ量カウンタ３０６の比率に基づいてＣＨＰの割り当てを行う（５２１）。すなわち、この場合、記憶制御装置１０は、ＣＨＰ０データ量カウンタ３０５およびＣＨＰ１データ量カウンタ３０６の値を比較して、値の小さい方、すなわち、その時点で処理負荷の小さいＣＨＰを当該フレームの処理用に割り当て、当該フレームの処理に必要なデータをそのＣＨＰに転送する。

【００２６】つぎに、（５０２）の処理において、フレームが起動先頭コマンドでなかった場合には、まず、フレーム情報（ＯＸ－ＩＤ，Ｓ－ＩＤ、ＬＰＮ番号、デバイス番号）をキーとした場合に該当するエクスチェンジをエクスチェンジ管理テーブル２００から検索する。そして、検索したエクスチェンジのＣＨＰ番号２０５に既に値がセットされていれば、当該フレームの処理に必要なデータをその値に対応するＣＨＰに送信する（５１６）。

【００２７】他方、ＣＨＰ番号２０５に値がセットされていなかった場合、すなわち、そのフレームが所属するエクスチェンジに既にＣＨＰが割り当てられていない場合には、（５０７）からの処理に進む。以上のようにして記憶制御装置１０は受信したフレームをつぎつぎに処理していくことになる。

【００２８】つぎに、以上の処理をより具体的に説明すべく、上位処理装置２０から記憶制御装置１０に対し、DX／LOC／WRCKDからなる１CCW（Channel Command Word）チェーン（例えば、「IBM 3990/9390 Storage Control Reference」を参照）のフレームが送られてきた場合の処理について、再度、図５のフローチャートに従って説明する。

【００２９】記憶制御装置１０は、DXコマンドを含んだフレームを受信すると、まず、このフレームが起動先頭コマンドフレームかどうかを判定する（５０８）。ここでDXコマンドはCCWチェーンの起動先頭コマンドであり、新たなエクスチェンジの起動先頭フレームであるので、記憶制御装置１０はこのフレームに対応するエクスチェンジを、エクスチェンジ管理テーブル２００に新規に登録する。

【００３０】つぎに記憶制御装置１０は、前記フレームのフレーム制御フィールド（F_ControlField）を参照し、当該フレームの後続フレームの存在有無を調べる（５０７）。その結果、後続フレームを有する場合には、さらにこのフレームがRD／WRいずれかのコマンドを含むフレームであるかどうかを調べる（５０８）。ここでDXコマンドは、READ／WRITEいずれのコマンドでも無いため、記憶制御装置１０は、当該フレームを通常のＣＨＰ割当方式により割り当てたＣＨＰにより処理することになる（５１０）。

【００３１】つぎに、記憶制御装置１０は、DXコマンドに引き続きCCWチェーンを構成するLOCコマンドを含むフレームを受信すると、このフレームが起動先頭コマンドであるかどうかを調べる（５０２）。ここでLOCコマンドが記載されたフレームは起動先頭コマンドで無いため、記憶制御装置１０はこのフレームのフレーム情報（ＯＸ－ＩＤ，Ｓ－ＩＤ、ＬＰＮ番号、デバイス番号）を用いてこのフレームに対応するエクスチェンジをエクスチェンジ管理テーブル２００より検索する。そして、この場合には、前述のDXコマンドが記載されたフレームにより前記フレーム情報に対応するエクスチェンジが既にエクスチェンジ管理テーブル２００に登録されているため、検索の結果、このエクスチェンジが検索されることになる。

【００３２】つぎに、記憶制御装置１０は、検索されたエクスチェンジについて、当該エクスチェンジの処理用に既にＣＨＰが割り当てられているかどうかを当該エクスチェンジのＣＨＰ番号２０５を参照して調べる（５０６）。ここでこのエクスチェンジには、まだＣＨＰの割り当てがされていないので、（５０７）の処理において当該フレームの後続チェーンが存在するかどうかを調べる。ここでこのフレームには後続チェーン（WRCKDコマンドのフレーム）が存在するため、（５０８）の処理が実行される。そして、LOCコマンドは、そのオペレーションコードからWRITEコマンドであることを判定できるコマンドであるため、（５１０）の処理へと移行して、当該LOCコマンドをWRデータキュー管理テーブル４１０に登録し、また、エクスチェンジ管理テーブル２００の当該フレームに対応するエクスチェンジのＣＨＰ番号２０５に「０」をセットする（５１１，５１３）。また、ＣＨＰ０データ量カウンタ３０５に当該LOCコマンドの処理されるデータ量、例えば、当該コマンドによりディスクユニットに書き込まれるデータのデータ量を加算する。

【００３３】つぎに、上位処理装置２０からWRCKDコマンドが記載されたフレームが送られてきた場合には、記

憶制御装置１０はこのフレームは起動先頭フレームでないため（５０２）、前述したLOCコマンドの場合と同様に当該フレームが所属するエクスチェンジ番号を検索する（５０５）。そして、この場合には、ＣＨＰ番号２０５に既に値がセットされており、当該フレームが所属するエクスチェンジに既にＣＨＰが割り当てられているため（５０６）、記憶制御装置１０は当該コマンドの処理に必要なデータをそのＣＨＰ番号に対応するＣＨＰに送信する。

【００３４】以上に説明したように、記憶制御装置１０はフレームに含まれるコマンドがRDコマンドであるかWRコマンドであるかに応じてそのフレームが所属するエクスチェンジの処理を行うＣＨＰを割り当てる。従って、ＣＨＰ０，１の双方が同時にWRもしくはRDの一方の処理を行う期間が減って、全二重方式の通信路の受信路５１と送信路５２の負荷のバランスが不均一になる期間を減らすことができる。

【００３５】ところで、以上のようなＣＨＰの割当方式を適用した場合でも、例えば、上位処理装置２０からWRコマンドもしくはRDコマンドのいずれか一方のみを含むフレームが連続して送信されて長期間一方のＣＨＰがその処理に占有されていたり、一のコマンドの処理対象となるデータのデータ量が大きい場合に前記の割当処理を実行してしまうと、かえって一方の通信路に負荷が片寄る結果となり、通信路の負荷のバランスが崩れてしまう可能性がある。そこで負荷分散をより徹底して行うようにするため、本発明の記憶制御装置１０はさらに以下に示すような各種の機能を備えている。

【００３６】このうち第１の機能は、ＣＨＰの割当処理を行うかどうかを、上位処理装置２０から送られてくるRDコマンドとWRコマンドのそれぞれにより処理されるデータ量の比率に応じて制御するようにしたものである。具体的には、図６に示すように、ポーリング処理などにより適宜なタイミングである一定期間におけるＣＨＰ割当処理管理テーブル３００の受信WRコマンドデータ量カウンタ３０２と、受信RDコマンドデータ量カウンタ３０３の増分からRDコマンドとWRコマンドのそれぞれにより処理されたデータ量の比率を算出（６０２）し、この比率がＣＨＰ割当処理管理テーブル３００のRD／WRデータ量比率閾値を超えた場合にはＣＨＰ割当処理管理テーブル３００のＣＨＰ割当処理実行フラグ３０１に「0」をセットし(６０５)、RD／WRデータ量比率閾値以下の場合にはＣＨＰ割当処理実行フラグ３０１に「1」をセットする（６０４）。なお、図６の例では、ポーリング機能などにより以上の処理を実行するたびに、受信WRデータカウンタと、受信RDデータカウンタの内容を初期化している（６０７）。

【００３７】第２の機能は、各ＣＨＰのキューの状態に応じてＣＨＰ割当処理実行フラグを制御するようにしたものである。具体的には、ポーリング処理などにより一定期間ごともしくは処理データが一定数に達する度ごとなどの適宜なタイミングで、RD／WRデータキュー管理テーブルから各ＣＨＰ０，１にキューイングされているコマンドの数およびこれらコマンドの処理対象となるデータの全データ量を算出する。キューイングされているコマンド数はRD／WRキュー管理テーブルのカウンタの値により把握される。また、全データ量は、データ転送量により把握される。RDコマンドおよびWRコマンドそれぞれのキューイング数の比率、RDコマンドおよびWRコマンドそれぞれのデータ転送量の比率を、それぞれＣＨＰ割当処理管理テーブル３００のＣＨＰキューデータ数比率閾値３０９、ＣＨＰキューデータ量比率閾値３１０と比較し、閾値を超えているかどうかに応じてＣＨＰ割当処理実行フラグ３０１を制御する。図７は以上の処理の一例を示すフローチャートである。

【００３８】第３の機能は、ＣＨＰ割当処理管理テーブル３００のＣＨＰ０，１データ量カウンタ３０５，３０６の比率に応じてＣＨＰ割当処理実行フラグ３０１を制御するようにしたものである。具体的には、一定時間毎にＣＨＰ０，１データ量カウンタ３０５，３０６の比率と、ＣＨＰ割当処理管理テーブル３００のＣＨＰデータ比率閾値３０７とを比較し、閾値を超えているかどうかに応じてＣＨＰ割当処理実行フラグ３０１を制御する。図８にこの場合の処理の一例を示す。

【００３９】第４の機能は、WRコマンドとRDコマンドのそれぞれについて、記憶制御装置１０が単位時間当たりに処理したデータ量（スループット）の比に応じてＣＨＰ割当処理実行フラグ３０１を制御するようにしたものである。ここでWRコマンドとRDコマンドそれぞれについてのスループットは、ＣＨＰ割当処理管理テーブル３００における受信WRコマンドデータ量カウンタ３０２および送信RDコマンドデータ量カウンタ３０４の変化率により算出する（９０１）。すなわち、図９に示すように、ある時刻におけるこれらカウンタ値とこれから単位時間経過後のこれらカウンタ値の差から単位時間当たりの処理データ量であるスループットを算出し（９０１～９０３，９０７，９０８）、このようにして求めたスループットの比がＣＨＰ割当処理管理テーブル３００のRD／WRスループット比率閾値３１１を超えているかどうかに応じてＣＨＰ割当処理実行フラグ３０４を制御する（９０４）。

【００４０】以上に説明した第１～第４の機能によれば、上位処理装置２０からWRコマンドもしくはRDコマンドのいずれか一方のみが記載されたフレームが連続して送られてきたり、一のコマンドの処理対象となるデータのデータ量が大きい場合における一方の通信路に負荷の片寄りを防ぐことが可能となり、フレームに含まれるコマンドの種類（RDコマンドもしくはWRコマンド）に応じてそのフレームを処理するＣＨＰを割り当てることで、全二重通信における受信路と送信路の負荷分散を図る前

述した仕組みをより一層効果的に機能させることが可能となる。なお、以上に説明した第１～第４の機能は、全てを一度に適用しなければならない訳ではなく、このうちの１の機能のみを適用したり、いくつかの機能を選択して適用するようにしてもよい。

【００４１】また、以上に説明した第１～第４の機能において、ポーリング処理の間隔を指定する数値、例えば、時間や処理データ数などの数値を、上位処理装置２０や、記憶制御装置１０に接続された運用管理端末などの外部装置から指定できるようにしてもよい。

【００４２】ところで、以上の実施例は、通信プロトコルがファイバチャネルプロトコルである場合について説明したが、上位処理装置２０と２以上のチャネルプロセサを備えた記憶制御装置１０が全二重方式で結ばれる構成を備えるデータ処理システムであれば、通信プロトコルの種類に限定されることなく適用することができる。

【００４３】以上の実施例は、記憶制御装置１０が一つのチャネル制御部１４に対し２つのＣＨＰ１３を備えている場合について説明したが、一つのチャネル制御部１４に対して３つ以上のＣＨＰ１３を備えている場合にも適用できることはもちろんである。

【００４４】

【発明の効果】以上に説明したように、本発明の記憶制御装置によれば、全二重方式の通信路を効率よく利用することができる。

【図面の簡単な説明】

【図１】本発明の一実施例によるデータ処理システムの概略構成を示す図である。

【図２】本発明の一実施例によるエクスチェンジ管理テーブルを示す図である。

【図３】本発明の一実施例によるＣＨＰ割当処理管理テーブルを示す図である。

【図４】本発明の一実施例によるRD／WRコマンドキュー管理テーブルを示す図である。

【図５】本発明の一実施例によるＣＨＰ割当処理を説明するフローチャートである。

【図６】本発明の一実施例による、ＣＨＰ割当処理を行うかどうかを、RDコマンドとWRコマンドのそれぞれにより処理されるデータ量の比率に応じて制御する処理を説明するフローチャートである。

【図７】本発明の一実施例による、ＣＨＰ割当処理を行うかどうかを、各ＣＨＰのキューの状態に応じてＣＨＰ割当処理実行フラグを制御する処理を説明するフローチャートである。

【図８】本発明の一実施例による、ＣＨＰ割当処理を行うかどうかを、ＣＨＰ０，１データ量カウンタの比率に応じて制御する処理を説明するフローチャートである。

【図９】本発明の一実施例による、ＣＨＰ割当処理を行うかどうかを、WRコマンドとRDコマンドのそれぞれについて、記憶制御装置が単位時間当たりに処理したデータ量（スループット）の比に応じて制御する処理を説明するフローチャートである。

【符号の説明】

１０　記憶制御装置
１１　キャッシュメモリ
１２　データ記憶手段
１３　チャネルプロセサ
１４　チャネル制御部
２０　上位処理装置
５１　受信路
５２　送信路

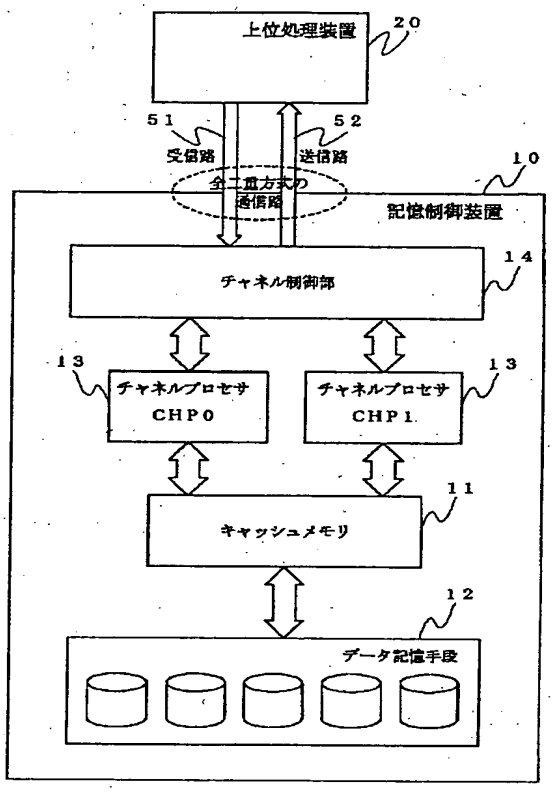
【図２】

201	202	203	204	205	206	207	208	209	210
EXG	ﾃｰﾌﾞﾙ有効ﾌﾗｸﾞ	ﾃﾞﾊﾞｲｽ番号	通信領域ﾎﾟｲﾝﾀ	CHP番号	論理ﾊﾟｽ	OX-ID	S-ID	D-ID	ﾃﾞｰﾀ量
1	1	05	0x0000	FF	0001	0004	0100	0200	2048
2	1	08	0x0040	01	0001	0005	0300	0400	4096
3	1	01	0x0080	00	0003	0006	0500	0600	8192
4	0								
:									
:									
X	0								

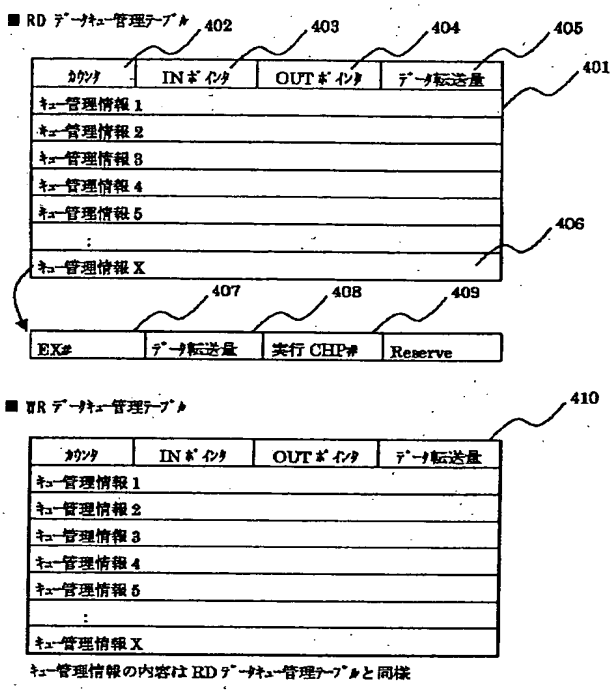
【図３】

CHP割当処理実行フラグ	301
受信WRコマンドデータ量カウンタ	302
受信RDコマンドデータ量カウンタ	303
送信RDコマンドデータ量カウンタ	304
CHP0データ量カウンタ	305
CHP1データ量カウンタ	306
CHPデータ比率閾値	307
RD/WRデータ比率閾値	308
CHPキューデータ数比率閾値	309
CHPキューデータ量比率閾値	310
RD/WRスループット比率閾値	311
モニタリング実行間隔時間	312
モニタリング実行間隔I/O数	313
モニタリング開始時刻	314
モニタリング終了時刻	315
I/O数カウンタ	316
ワークエリア	317

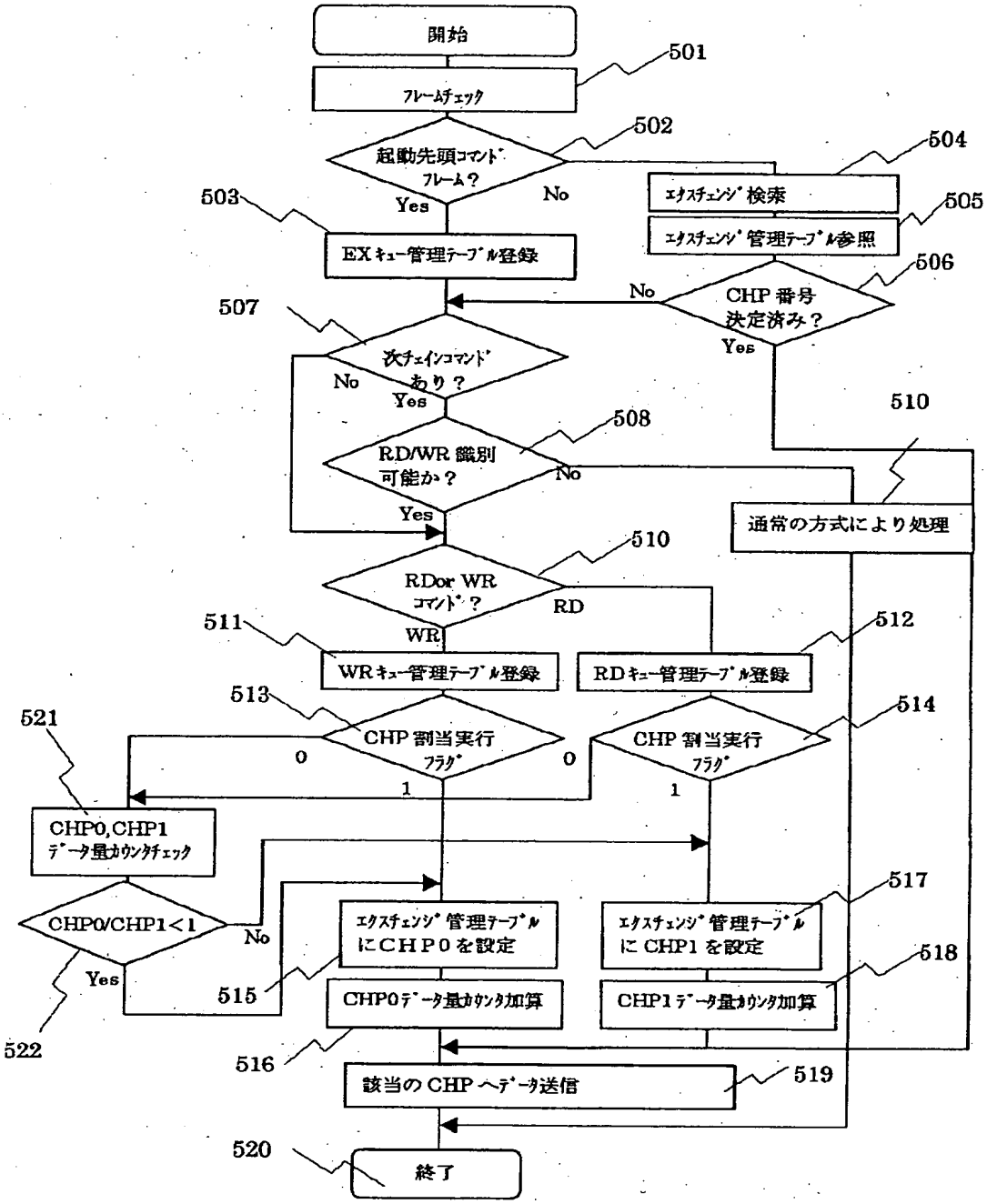
【図１】


上位処理装置
20
51
受信路
52
送信路
全二重方式の
通信路
10
記憶制御装置
14
チャネル制御部
13
チャネルプロセサ
CHP0
13
チャネルプロセサ
CHP1
11
キャッシュメモリ
12
データ記憶手段

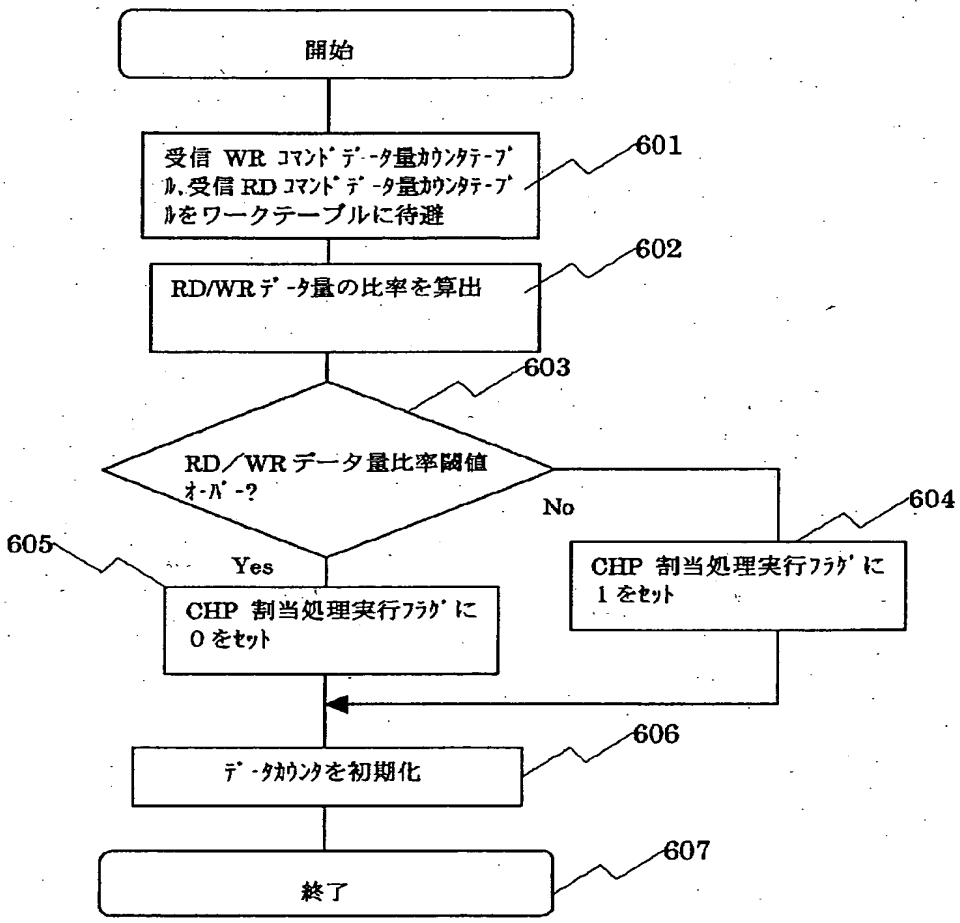
【図４】


■ RD データキュー管理テーブル
402
403
404
405
401
カウンタ
INポインタ
OUTポインタ
データ転送量
キュー管理情報 1
キュー管理情報 2
キュー管理情報 3
キュー管理情報 4
キュー管理情報 5
キュー管理情報 X
406
407
408
409
EX#
データ転送量
実行 CHP#
Reserve
■ WR データキュー管理テーブル
410
カウンタ
INポインタ
OUTポインタ
データ転送量
キュー管理情報 1
キュー管理情報 2
キュー管理情報 3
キュー管理情報 4
キュー管理情報 5
キュー管理情報 X
キュー管理情報の内容は RD データキュー管理テーブルと同様

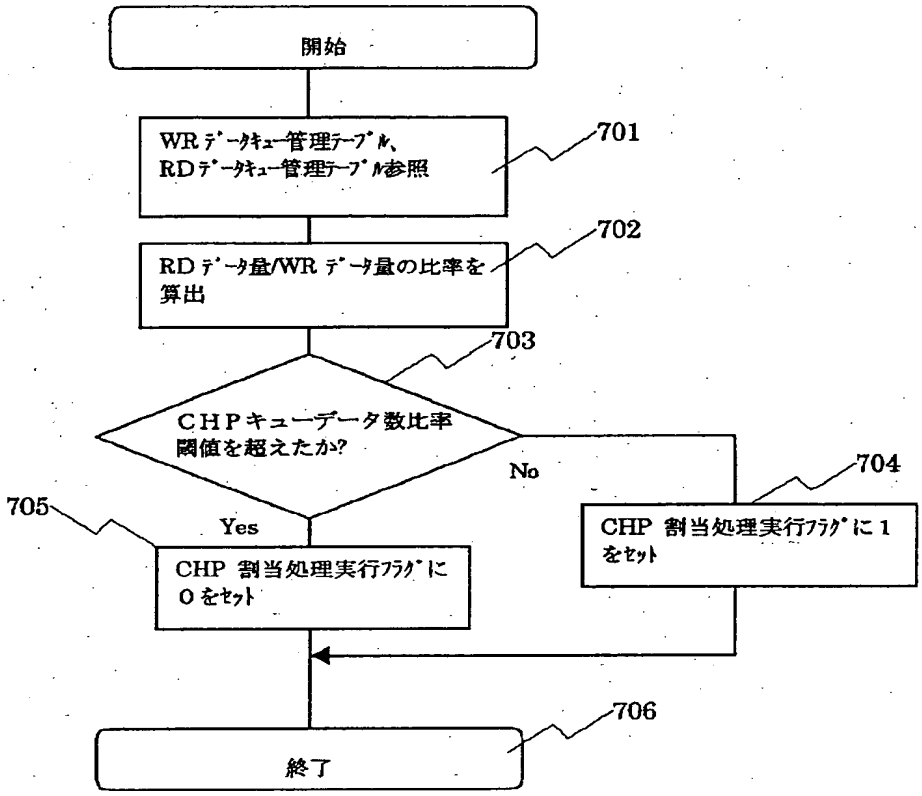
【図５】


開始
501
フレームチェック
502
起動先頭コマンド フレーム？
Yes
No
503
EXキュー管理テーブル登録
504
エクスチェンジ検索
505
エクスチェンジ管理テーブル参照
506
CHP番号 決定済み？
No
Yes
507
次チェインコマンド あり？
No
Yes
508
RD/WR 識別 可能か？
No
Yes
510
通常の方式により処理
510
RDor WR コマンド？
RD
WR
511
WRキュー管理テーブル登録
512
RDキュー管理テーブル登録
513
CHP 割当実行 フラグ
0
1
514
CHP 割当実行 フラグ
0
1
521
CHP0,CHP1 データ量カウンタチェック
CHP0/CHP1＜1
No
Yes
522
515
エクスチェンジ管理テーブル にCHP0を設定
CHP0データ量カウンタ加算
517
エクスチェンジ管理テーブル にCHP1を設定
518
CHP1データ量カウンタ加算
516
該当のCHPへデータ送信
519
520
終了

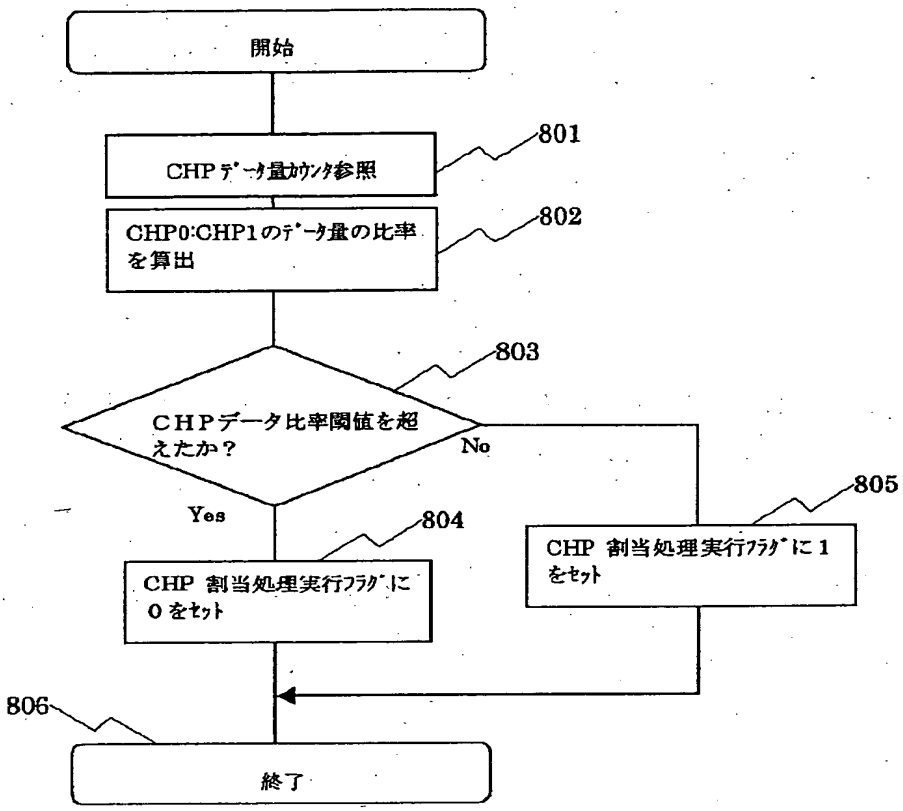
【図６】


開始
受信 WR ｺﾏﾝﾄﾞﾃﾞｰﾀ量ｶｳﾝﾀﾃｰﾌﾞﾙ、受信 RD ｺﾏﾝﾄﾞﾃﾞｰﾀ量ｶｳﾝﾀﾃｰﾌﾞﾙをワークテーブルに待避
601
RD/WR ﾃﾞｰﾀ量の比率を算出
602
603
RD／WR データ量比率閾値ｵｰﾊﾞｰ?
No
Yes
604
CHP 割当処理実行ﾌﾗｸﾞに１をｾｯﾄ
605
CHP 割当処理実行ﾌﾗｸﾞに０をｾｯﾄ
606
ﾃﾞｰﾀｶｳﾝﾀを初期化
607
終了

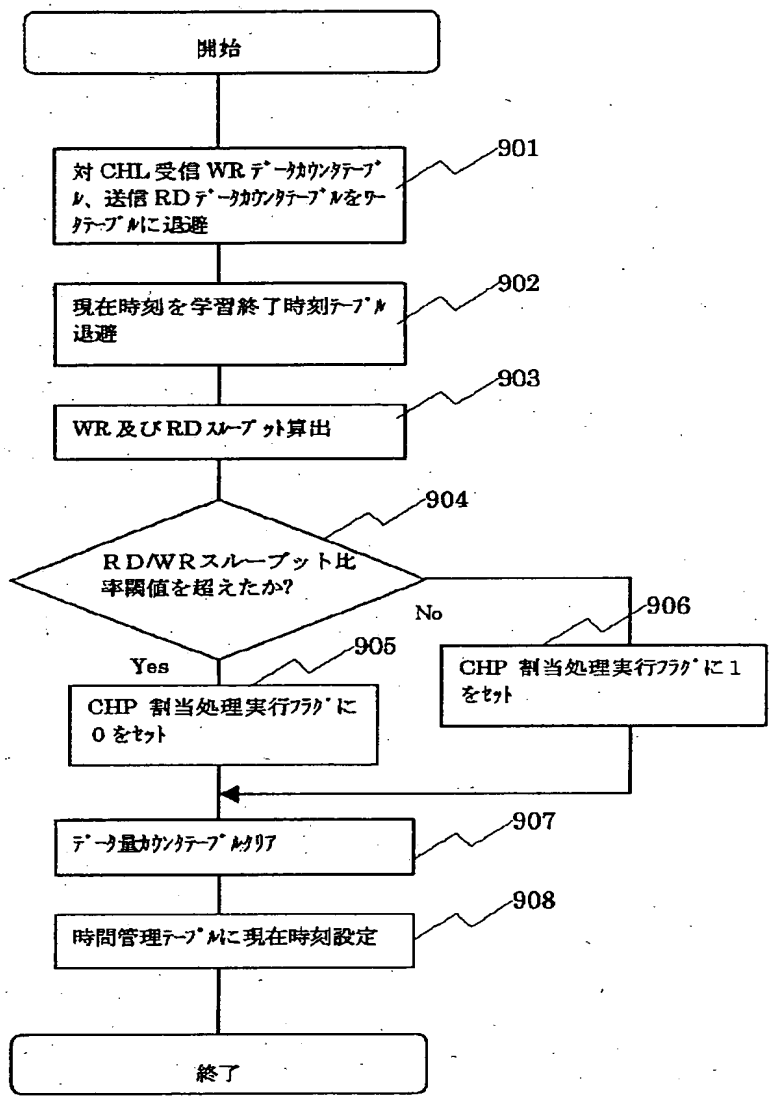
【図７】


開始
WR ﾃﾞｰﾀｷｭｰ管理ﾃｰﾌﾞﾙ、
RD ﾃﾞｰﾀｷｭｰ管理ﾃｰﾌﾞﾙ参照
701
RD ﾃﾞｰﾀ量/WR ﾃﾞｰﾀ量の比率を
算出
702
703
ＣＨＰキューデータ数比率
閾値を超えたか?
No
704
CHP 割当処理実行ﾌﾗｸﾞに１
をｾｯﾄ
705
Yes
CHP 割当処理実行ﾌﾗｸﾞに
０をｾｯﾄ
706
終了

【図８】


開始
CHPデータ量カウンタ参照
801
CHP0:CHP1のデータ量の比率を算出
802
CHPデータ比率閾値を超えたか？
803
No
Yes
804
CHP 割当処理実行フラグに0をセット
805
CHP 割当処理実行フラグに1をセット
806
終了

【図９】



フロントページの続き

(72)発明者　榊　豪紀
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内

(72)発明者　塚田　大
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内

Ｆターム(参考)　5B014 EB04 FB02 GD04 GD34
5B065 BA01 CA02 CA11 CC08 ZA13

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-222339
(43)Date of publication of application : 11.08.2000

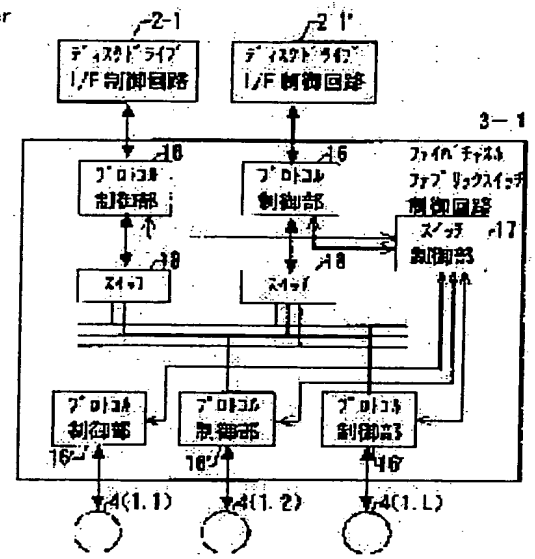
(51)Int.Cl. G06F 13/14
G06F 3/06
G06F 13/00
G06F 13/10

(21)Application number : 11-024648
(22)Date of filing : 02.02.1999

(71)Applicant : HITACHI LTD
(72)Inventor : ARIGA KAZUHISA

(54) DISK SUB-SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To connect plural disk drives with a disk drive interface circuit without scarifying transmitting performance by using a fiber channel fabric topology for reducing the number of connection lines by using a fiber channel interface being a serial interface, and for realizing switch connection.
SOLUTION: A fiber channel fabric switch control circuit 3 is provided between a disk drive 4 and a disk drive interface control circuit 2, and protocol control part 16 is provided between a switch 18 in the fabric switch circuit 3 and the disk drive.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－222339
（P2000－222339A）
(43)公開日　平成12年８月11日(2000.8.11)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'（参考）
Ｇ０６Ｆ　13/14	３１０	Ｇ０６Ｆ　13/14	３１０Ｆ	５Ｂ０１４
3/06	３０１	3/06	３０１Ａ	５Ｂ０６５
	３０５		３０５Ｃ	５Ｂ０８３
	５４０		５４０	
13/00	３０１	13/00	３０１Ｄ	

審査請求　未請求　請求項の数５　ＯＬ　（全 10 頁）　最終頁に続く

(21)出願番号　特願平11－24648

(22)出願日　平成11年２月２日(1999.2.2)

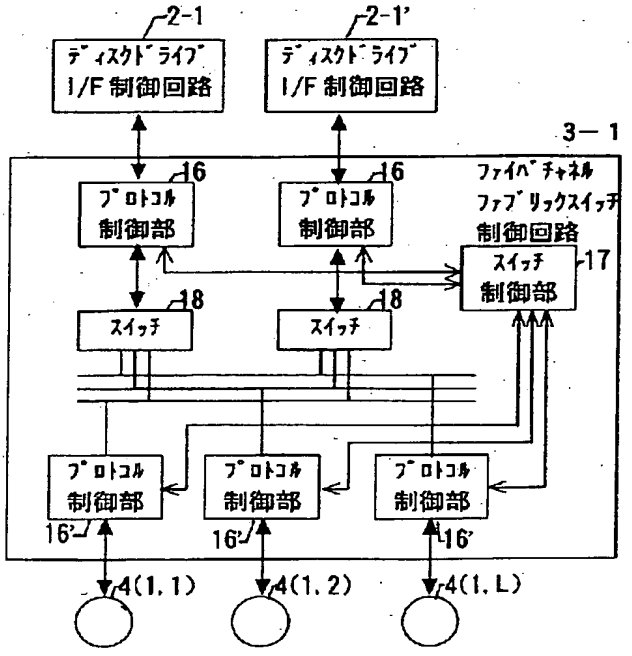
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　有賀　和久
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内
(74)代理人　100068504
弁理士　小川　勝男

Ｆターム(参考)　5B014 EA02 EA04 EB05 HA09 HA12
5B065 BA01 CA11 CA19 CA30 CE12
EA25 ZA11
5B083 AA08 BB01 BB03 CC02 CD13
EE08 EF11

(54)【発明の名称】　ディスクサブシステム

(57)【要約】
【課題】ディスク制御装置とディスクドライブの接続においては、ＳＣＳＩを用いたインタフェースが主流であるが、ディスクドライブが増加した場合に１本のインタフェースで１対１の接続を行う場合、現状のファイバチャネルを用いたディスクドライブでは、スイッチ接続が出来ない形式となっているので多数のインタフェースが必要となり、実装面で困難が生ずる。
【解決手段】ディスクドライブ４とディスクドライブインタフェース制御回路２との間にファイバチャネル・ファブリック・スイッチ制御回路３を設け、このファイバチャネル・ファブリック・スイッチ回路３内のスイッチ１８とディスクドライブとの間にプロトコル制御部１６を設ける。

図 3



【特許請求の範囲】
【請求項１】データを記憶する複数のディスクドライブと、このディスクドライブ及びホストコンピュータからのデータの入出力を制御するディスクアレイ制御部とを有し、このディスクアレイ制御部と前記ディスクドライブとをファイバチャネルにて接続したディスクサブシステムにおいて、
前記ディスクアレイ制御部と前記ディスクドライブとをスイッチ接続したディスクアレイシステム。
【請求項２】データを記憶する複数のディスクドライブと、このディスクドライブ及びホストコンピュータからのデータの入出力を制御するディスクアレイ制御部とを有するディスクサブシステムにおいて、
前記ディスクドライブと前記ディスクアレイ制御部との間にスイッチとこのスイッチの切換え制御をするスイッチ制御部を設け、前記スイッチと前記ディスクドライブとの間、及び／または前記ディスクアレイ制御部と前記スイッチとの間にプロトコル制御部を設けたディスクサブシステム。
【請求項３】前記ディスクアレイ制御部と前記スイッチとの間、及び前記スイッチと前記ディスクドライブとの間とはファイバチャネルを用いて接続したものであり、前記スイッチはファイバチャネルファブリックスイッチである請求項２に記載のディスクサブシステム。
【請求項４】ホストコンピュータからのデータの入出力を制御するホストインタフェース制御部と、このホストインタフェース制御部で受けたデータを一時的に格納するキャッシュメモリと、前記データにパリティデータを付加するパリティデータ生成部と、前記データ及び前記パリティデータを記憶する複数のディスクドライブと、このディスクドライブに前記データを書き込むディスクドライブインターフェイスとからなるディスクアレイ制御部とを有するディスクサブシステムにおいて、
前記ディスクドライブインターフェイスにプロトコル制御部とスイッチを設け、前記複数のディスクドライブをスイッチ接続したディスグサブシステム。
【請求項５】ホストコンピュータからのデータの入出力を制御するホストインタフェース制御部と、このホストインタフェース制御部で受けたデータを一時的に格納するキャッシュメモリと、前記データにパリティデータを付加するパリティデータ生成部と、前記データ及び前記パリティデータを記憶する複数のディスクドライブと、このディスクドライブに前記データを書き込むディスクドライブインターフェイスとからなるディスクアレイ制御部とを有するディスクサブシステムにおいて、
前記ディスクアレイ制御部と前記ディスクドライブとの間をファイバチャネルを用いて接続し、前記ディスクアレイ部と前記ディスクドライブとの間に、アクセス対象となる前記ディスクドライブのＩＤ番号検出及びファイバチャネル・プロトコルの制御を行い前記ディスクドライブインタフェースと接続される第一のプロトコル制御部と、各ディスクドライブのＩＤ番号を記憶しておりＩＤ番号によりスイッチを設定するスイッチ制御部と、前記ディスクドライブにこのＩＤ番号を割り付ける前記ディスクドライブ４と接続される第二のプロトコル制御とを備えたファブリックスイッチを設けたディスクサブシステム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、ディスクサブシステム、ディスクアレイ、ディスクドライブを内蔵した計算機等の電子機器に関し、特にアレイディスクをファブリックスイッチ接続し高速転送を可能とする技術に関する。
【０００２】
【従来の技術】一般に、ディスクアレイにおいてディスク制御装置と複数のディスクドライブとを接続する場合には、特開平１０－１７１７４６号公報に記載のようにＳＣＳＩインタフェース、若しくはファイバチャネル・アービトレイテッドループ・トポロジが利用されている。
【０００３】ＳＣＳＩインタフェースは、同一線路上にデータを時分割して転送する方式をとっており、イニシエータに対するアクセスは、１伝送路上に１時刻あたり１対１の通信を行う方式である。
【０００４】ファイバチャネル・アービトレイテッドループ・トポロジでは、ＳＣＳＩインタフェースに対して、シリアルインタフェースによりループ状にイニシエータ、ディスクドライブを接続することができ、フレームに分割されたデータを時分割して転送し、同時に多数デバイスの通信が行え、接続可能ディスクドライブ数も１２６と拡張できる。
【０００５】
【発明が解決しようとする課題】今後ディスクドライブの小型化・高密度化により、より多くのディスクドライブを使用することが可能となると考えられる。
【０００６】ＳＣＳＩインタフェースは、１伝送路上に１時刻あたり１対１の通信を行う方式であるため同時に多数のイニシエータとディスクドライブの通信ができない。また、接続可能なディスクドライブの数も７から１５台と少ない。そのためＳＣＳＩを用いたインタフェースでドライブが増加した場合に１本のインタフェースで１対１の接続を行おうとすると、多数のインタフェースが必要となり、実装面で困難が生ずる。また、１つの制御回路で接続可能なディスクドライブの数が少ないため、多数の制御回路を使用する必要が生じる。
【０００７】一方、ファイバチャネルを使用した場合、ディスクドライブはプロトコルが制御装置とは異なるためにスイッチ接続が出来ず、多数のディスクドライブが同一ループを共有するファイバチャネル・アービトレイ

テッドループを用いてループ接続とせざるを得なかった。そのため、同一ループに接続されるディスクドライブ数を増加すると、ディスクドライブのデータ転送速度がループの最大データ転送速度よりも大きくなり、結果的にループの最大データ転送速度以上の効率では転送が行えなくなりＳＣＳＩインタフェースと同程度のデータ転送速度でしか接続できなかった。

【０００８】

【課題を解決するための手段】上記課題を解決するため本発明では、ディスクドライブと制御装置とをスイッチ接続を可能とするため、プロトコル制御部をファイバチャネル・ファブリック・スイッチとディスクドライブとの間に設ける。

【０００９】

【発明の実施の形態】以下、図面を用いて本発明を適用した外部記憶装置（ディスクサブシステム）の実施例を説明する。図１は全体図である。

【００１０】図に示す外部記憶装置において、Ｎ個のディスクアレイ制御回路（制御部）（１－１）～（１－Ｎ）（途中の１－２等は省略、以下同じ）は、上位側はホストコンピュータ（図示せず）に接続され、下位側はＭ個のディスクドライブインターフェイス（ディスクドライブＩ／Ｆ）制御回路（２－１）～（２－Ｍ）を備えている。ディスクアレイ制御回路のハード構成の詳細は後述する。Ｍ個のファイバチャネル・ファブリック・スイッチ制御回路（３－１）～（３－Ｍ）は、ファイバチャネル・インタフェース５によってディスクドライブを制御するディスクドライブインターフェイス（Ｉ／Ｆ）制御回路（２－１）～（２－Ｍ）にそれぞれ接続されている。そして一つのファイバチャネル・ファブリック・スイッチ制御回路に対してＬ個、計Ｍ×Ｌ個のディスクドライブ（４（１，１）～４（Ｍ，Ｌ））は、ファイバチャネル・インタフェース６によって、ファイバチャネル・ファブリック・スイッチ制御回路（３－１）～（３－Ｍ）と接続されている。

【００１１】また、各ディスクドライブインタフェース制御回路（２－１）～（２－Ｍ）、及びデータを格納しておくディスクドライブ４（１，１）～４（Ｍ，Ｌ）はそれぞれ個別の識別子（ＩＤ番号）をもつ。ファイバチャネル・ファブリック・スイッチ制御回路（３－１）～（３－Ｍ）は、ディスクドライブインタフェース制御回路（２－１）～（２－Ｍ）から接続するディスクドライブのＩＤ番号を受け取り、対応するディスクドライブインタフェース制御回路（２－１）～（２－Ｍ）とディスクドライブ４（１，１）～４（Ｍ，Ｌ）の１対１の接続を確立する。

【００１２】図２にディスクアレイ制御回路（１－１）～（１－Ｎ）のハードウェア構成を示す。上位ホストコンピュータ（図示せず）から転送されるデータは、ホストインタフェース制御部７により制御されキャッシュメモリ８に一時格納されると共にパリティデータ生成部９によりパリティデータを付加され、データブロックとパリティデータブロックとに分解（全体でＭ個）される。これらのデータ及びパリティのブロックは、それぞれ対応するインタフェースであるディスクドライブインタフェース制御回路（２－１）～（２－Ｍ）によりディスクドライブグループ（図示せず）に格納される。

【００１３】上位ホストコンピュータにデータを転送する場合は、転送するデータがキャッシュメモリ８に存在する場合には、そのデータをホストインタフェース制御部７が上位ホストコンピュータに転送する。転送するデータがキャッシュメモリ８に存在しない場合には、ディスクドライブインタフェース制御回路（２－１）～（２－Ｍ）がディスクドライブグループより分解されたデータを読み出し、パリティデータ生成部９で分解されたデータを結合した後にキャッシュメモリ８に一時格納するとともにホストインタフェース制御部７が上位ホストコンピュータに転送する。

【００１４】なお、以上の例はＲＡＩＤを用いた場合のデータ格納方法であり、ＲＡＩＤ方式を用いずにデータを格納することも当然可能である。その場合にはパリティデータ生成部９が存在せずに上位ホストコンピュータ（図示せず）から転送されるデータは、ホストインタフェース制御部７によりキャッシュメモリ８に一時格納されると共にディスクドライブグループ内の何れかのディスクドライブに格納され、ミラー方式の場合には、複数のディスクドライブに同一のデータを複数格納する。読み出す際にもディスクドライブからデータを読み出し、キャッシュメモリ８に一時格納するとともにホストインタフェース制御部７が上位ホストコンピュータに転送する。

【００１５】以下の例もＲＡＩＤを使用したディスクサブシステムについて説明するが、ＲＡＩＤを用いた場合に限らないことはもちろんである。

【００１６】図３にファイバチャネル・ファブリック・スイッチ制御回路（３－１）～（３－Ｍ）のハードウェア構成を示す。ディスクドライブインタフェース制御回路（２－１）と接続されるプロトコル制御部１６（第一のプロトコル制御部）は、アクセス対象となるディスクドライブ４（１，１）～４（１，Ｌ）のＩＤ番号検出及びファイバチャネル・プロトコルの制御を行う。ディスクドライブ４（１，１）～４（１，Ｌ）と接続されるプロトコル制御部１６’（第二のプロトコル制御部）はディスクドライブ４（１，１）～４（１，Ｌ）にＩＤ番号を割り付け、スイッチ制御部１７に担当するディスクドライブ４（１，１）～４（１，Ｌ）のＩＤ番号を報告する。スイッチ制御部１７は、各ディスクドライブ４（１，１）～４（１，Ｌ）のＩＤ番号を記憶しており、ディスクドライブインタフェース制御回路（２－１）～（２－Ｍ）より受領したＩＤ番号によりスイッチ１８を

設定し、１対１の接続を確立する。

【００１７】尚、プロトコル制御はプロトコル制御部１６’側で行うように設定してもよいし、ホストコンピュータからのデータ転送時とホストコンピュータへデータ転送時とや、通常のデータ転送とディスク障害時のデータ移送とでプロトコル制御部１６とプロトコル制御部１６’とを切り換えるように設定してもよい。

【００１８】また、プロトコル制御部１６或いはプロトコル制御部１６’の何れか一方のみとし、プロトコル制御部１６の代わりにＩＤ番号検出手段を設ける、或いはプロトコル制御部１６’の代わりにＩＤ番号割り付け手段を設けてもよい。

【００１９】また、ファイバチャネル・ファブリック・スイッチを独立した装置としてではなく、ディスクドライブインターフェイス制御回路（２－１）～（２－Ｍ）内にプロトコル制御部とスイッチとを設け、直にディスクドライブ４（１，１）～４（１，Ｌ）と接続するようにしてもよい。

【００２０】図４にファイバチャネルファブリックスイッチ制御回路（３－１）～（３－Ｍ）の動作を示す。

【００２１】ディスクアレイ制御回路（１－１）は、Ｍ個に分解されたデータをディスクドライブグループ（１０－１）に格納する。この際、ディスクアレイ制御回路（１－１）のＭ個のディスクドライブインタフェース制御回路（２－１）～（２－Ｍ）は、ファイバチャネル・ファブリック・スイッチ制御回路（３－１）～（３－Ｍ）に対し、ディスクドライブグループ（１０－１）に属するディスクドライブのＩＤ番号を送信し、スイッチの確立を行う。ファイバチャネル・ファブリック・スイッチ制御回路（３－１）～（３－Ｍ）内のプロトコル制御部１６（図３参照）は、ＩＤ番号を検出し、スイッチ制御部１７にスイッチ接続の切替を要求する。そしてディスクドライブに合わせたプロトコル制御を行う。スイッチ制御部１７（図３参照）はスイッチ１８（図３参照）を接続要求もとのディスクアレイ制御回路（１－１）と接続要求先のディスクドライブグループ（１０－１）に属するディスクドライブ４とを接続するよう切り替える。

【００２２】このとき、ディスクアレイ制御回路（１－１）は、ディスクドライブグループ（１０－１）とファイバチャネル・ファブリック・スイッチ制御回路（３－１）～（３－Ｍ）を介して１対１で対応しているので、他のディスクアレイ制御回路（１－Ｎ）と他のディスクドライブグループ（１０－２）は独立して他のデータ転送を行うことが出来る。つまり、ディスクアレイ制御回路（１－Ｎ）がディスクドライブグループ（１０－Ｌ）に対する接続の確立を行っても、ディスクアレイ制御回路（１－１）とディスクドライブグループ（１０－１）及びディスクアレイ制御回路（１－Ｎ）とディスクドライブグループ（１０－Ｌ）との接続は互いに独立して動作することができるので、それぞれのディスクアレイ制御回路及びディスクドライブ間で可能となる最高のデータ転送速度でデータ転送を行うことができる。

【００２３】尚、詳細は説明しないが、スイッチ制御部１７は上記のスイッチ切換えを行うと共に、データ読み書きの際にディスクドライブが既に読み書きを出来る状態になったという信号を受けてスイッチ１８の接続切換えを行うことで転送時間を有効に最大限確保することができる。

【００２４】図５に本発明の拡張された実施例を示す。先に示した実施例において、ファイバチャネル・ファブリック・スイッチ制御回路３のプロトコル制御部１６とディスクドライブ４とを１対１で対応させて接続していた部分を、プロトコル制御部１６からファイバチャネル・アービトレイテッド・ループ制御回路１１を介して複数のディスクドライブ４をループ接続するしている。この様に接続することで、安価なディスクドライブ４を多数の接続することで大容量のディスクドライブを備えた場合と同等な性能にできる。この場合でも、全てのディスクドライブがループ接続となる訳ではなく、見かけ上はファイバチャネル・アービトレイテッド・ループ制御回路１１と多数のディスクドライブ４で一つのディスクドライブ４であるので、アクセス性能は低下することがない。

【００２５】また図示はしないが、ディスクドライブのアクセス速度に対し、ファイバチャネルインタフェースの最大データ転送速度に充分余裕がある場合には、複数のディスクドライブ４をファイバチャネル・アービトレイテッド・ループ制御回路１１に接続し、複数のディスクドライブを同一ループ内に接続し、ファイバチャネルの最大転送レートを複数のディスクドライブ４で共有することで、アクセス性能を低下させることなくディスクドライブ４を増加させることも可能である。

【００２６】図６に図５に示した実施例に用いるアービトレイテッドループ制御回路１１のハードウェア構成図を示す。

【００２７】アービトレイテッドループ制御回路１１は、ループバイパス回路１３と複数のディスクドライブ接続ポート１２、及びファブリックスイッチ接続ポート１５からなる。ディスクドライブ４からはループバイパス回路切替信号１４が出力され、ディスクドライブ障害時にはポートをバイパスさせ、ループを切断することなく、他の動作しているディスクドライブへ影響を与えずにディスクドライブの取り外し、追加を行うことを可能とする。

【００２８】図７に本発明の他の拡張された実施例を示す。

【００２９】本実施例は、各ファイバチャネル・ファブリック・スイッチ制御回路（３－１）～（３－Ｍ）に接続されるスペアディスク制御回路１９と、このスペアデ

ィスク制御回路１９に接続される複数のスペアディスクドライブ（４－ａ），（４－ｂ）を備えている。そしてファイバチャネル・ファブリック・スイッチ制御回路３内では、故障したディスクドライブ４を含むディスクドライブグループ（図では４（１，２）のディスクドライブグループ）と接続しているプロトコル制御部１６’（図３参照）は、スイッチ１８を介してスペアディスク制御回路１９と接続しているプロトコル制御部１６’に接続される。何れかのディスクドライブが障害を起こした場合、ディスクアレイ制御回路（１－１）～（１－Ｎ）は、スペアディスクドライブ（４－ａ）または（４－ｂ）にデータの再構築を行う。

【００３０】特定のディスクドライブ４にエラーが多発し故障のおそれが出た場合には、エラーが多発するディスクドライブ４のデータをスペアディスクドライブ（４－ａ）または（４－ｂ）に移管させ再構築を行う。ディスクドライブ４が完全に破損してしまいデータの移管が不可能な場合には、破損したディスクドライブ４のディスクドライブグループのデータを用いて、図２に示したキャッシュメモリ８とパリティデータ生成部９にて破損データを再生しスペアディスクドライブ（４－ａ）または（４－ｂ）に書き込む。

【００３１】或いは、スペアディスク制御回路１９が独立して行うようにしてもよい。そのためこのスペアディスク制御回路１９内にキャッシュメモリやパリティデータ再生部を備える。そして、ディスクドライブ４が完全に破損した場合には、残りのディスクドライブグループのデータをスペアディスク制御回路１９で読み込み、破損データを再生してスペアディスクドライブ（４－ａ）または（４－ｂ）に書き込むようにする。

【００３２】そのため、故障したディスクドライブ４或いは故障個所を修復するためにパリティデータを含め分割されたデータを記憶した各ディスクドライブからスペアディスク制御回路１９へのアクセスと、ディスクアレイ制御回路（１－１）～（１－Ｎ）を介して行うディスクドライブ４（図では４（１，１）及び４（１，Ｌ）のディスクドライブグループ）とホストコンピュータからのデータアクセスとが独立して動作可能となることで、ホストコンピュータのデータアクセスに影響を与えずにデータの再構築を行うことを可能とする。

【００３３】また、障害ディスクドライブが正常ディスクドライブと取り替えられた場合も同様にして、スペアディスク制御回路１５がファイバチャネル・ファブリック・スイッチ制御回路（３－１）～（３－Ｍ）に対し、スペアディスグドライブ（４－ａ），（４－ｂ）と、障害ディスクドライブから取り替えられた正常ディスクドライブと１対１の接続を行い、ディスクアレイ制御回路（１－１）～（１－Ｎ）とディスクドライブグループ（１０－１）～（１０－Ｌ）（図４参照）とのアクセスを妨げることなく、独立してデータのコピーを行うことで、ホストコンピュータからのアクセスにまったく影響なく障害ディスクドライブの復旧を行うことができる。

【００３４】

【発明の効果】本発明により、シリアルインタフェースであるファイバチャネルインタフェースを用い接続線数を減少させ、さらにスイッチ接続を可能とするファイバチャネル・ファブリック・トポロジを用いることでディスクドライブインタフェース回路に多数ディスクドライブを伝送性能を犠牲にすることなく接続することが可能となる。また、各制御装置、ディスクドライブグループ毎に接続を動的に切り替えることで、少数のディスクドライブ制御回路で多数のディスクドライブを制御することができる。更に、ディスクドライブ障害時のデータ移行をディスクドライブインタフェース制御回路とディスクドライブのデータ転送と独立して行うことでシステムの信頼性を向上させることができる。

【図面の簡単な説明】

【図１】本発明実施例の全体図である。

【図２】ディスクアレイ制御回路の詳細図である。

【図３】ファイバチャネルファブリックスイッチ制御回路の詳細図である。

【図４】ファイバチャネルファブリックスイッチの接続図である。

【図５】ファイバチャネルファブリックスイッチとアービトレイテッドループの接続図である。

【図６】ファイバチャネルアービトレイテッドループ制御回路の詳細図である。

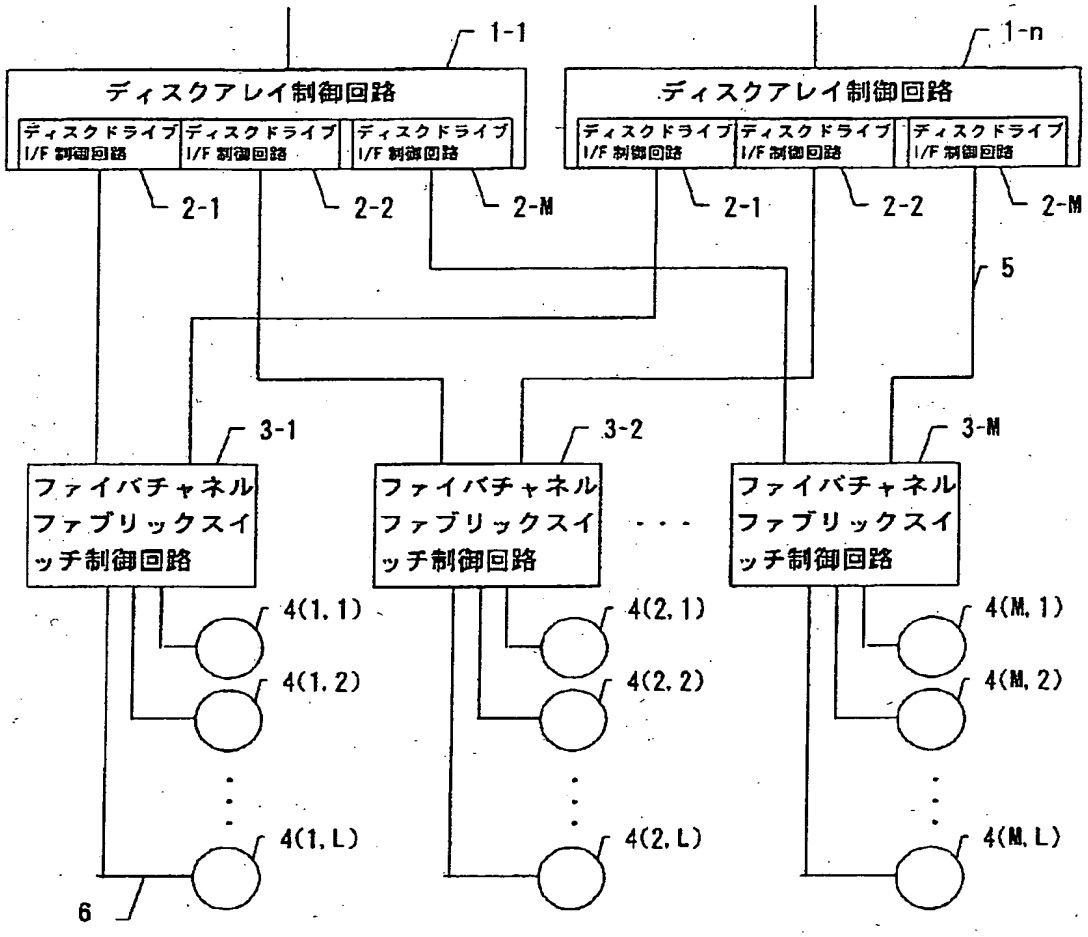
【図７】スペアディスク制御回路の接続図である。

【符号の説明】

１…ディスクアレイ制御回路、２…ディスクドライブインタフェース制御回路、３…ファイバチャネルファブリックスイッチ制御回路、４…ディスクドライブ、５…ファイバチャネルインタフェース、６…ファイバチャネルインタフェース、７…ホストインタフェース制御部、８…キャッシュメモリ、９…パリティデータ生成部、１０…ディスクドライブグループ、１１…ファイバチャネルアービトレイテッドループ制御回路、１２…ディスクドライブ接続ポート、１３…ループバイパス回路、１４…ループバイパス信号切替信号、１５…ファブリックスイッチ接続ポート、１６…プロトコル制御部、１７…スイッチ制御部、１８…スイッチ、１９…スペアディスク制御回路。

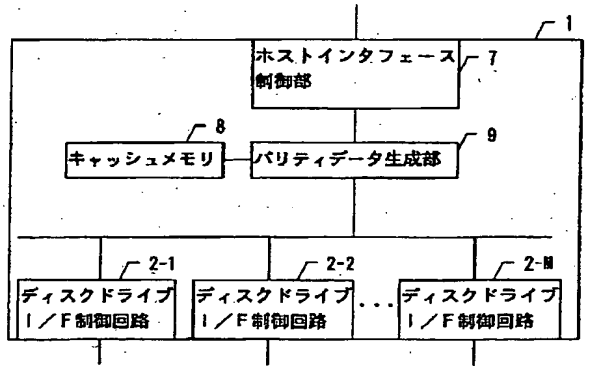
【図１】

図 1



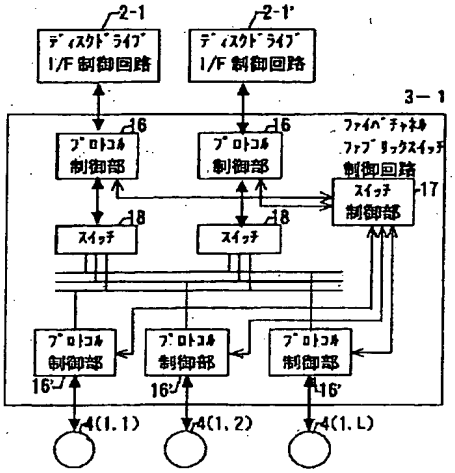
【図２】

図 2



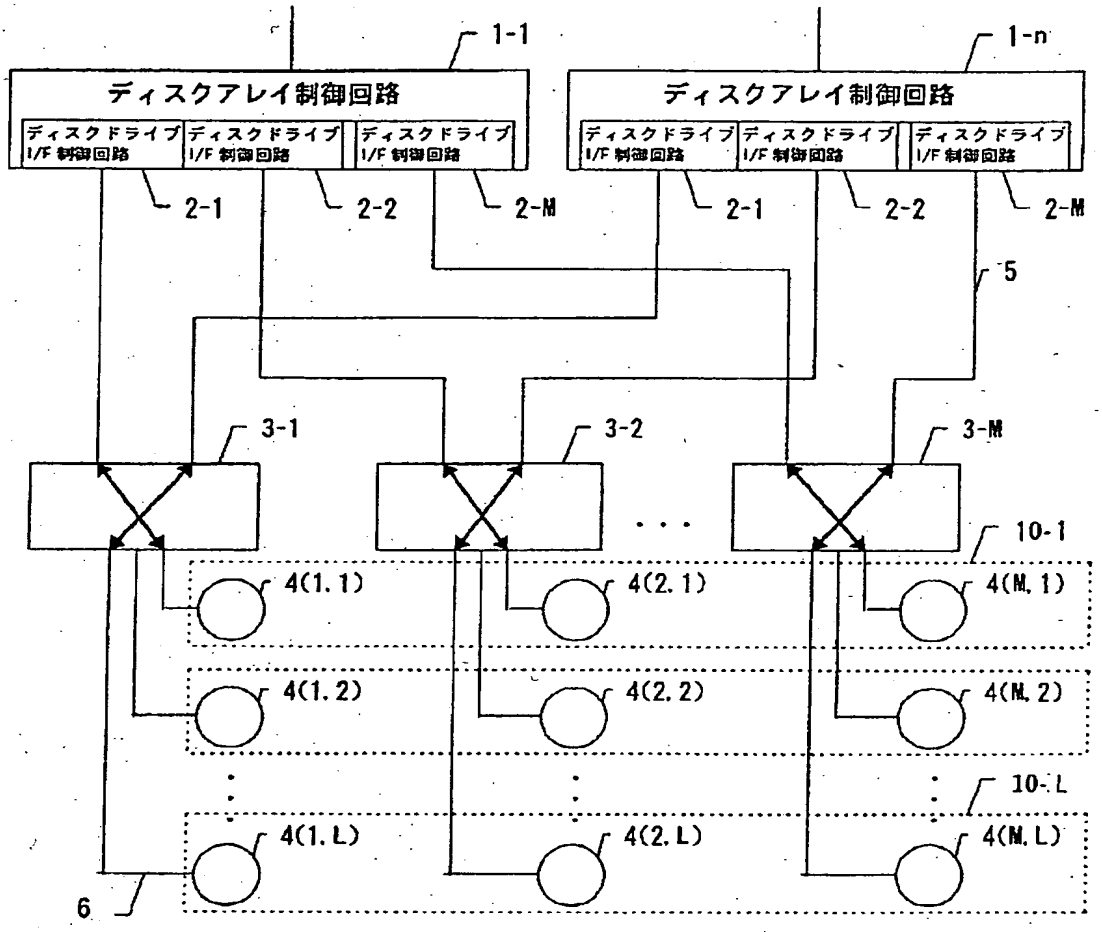
【図３】

図 3



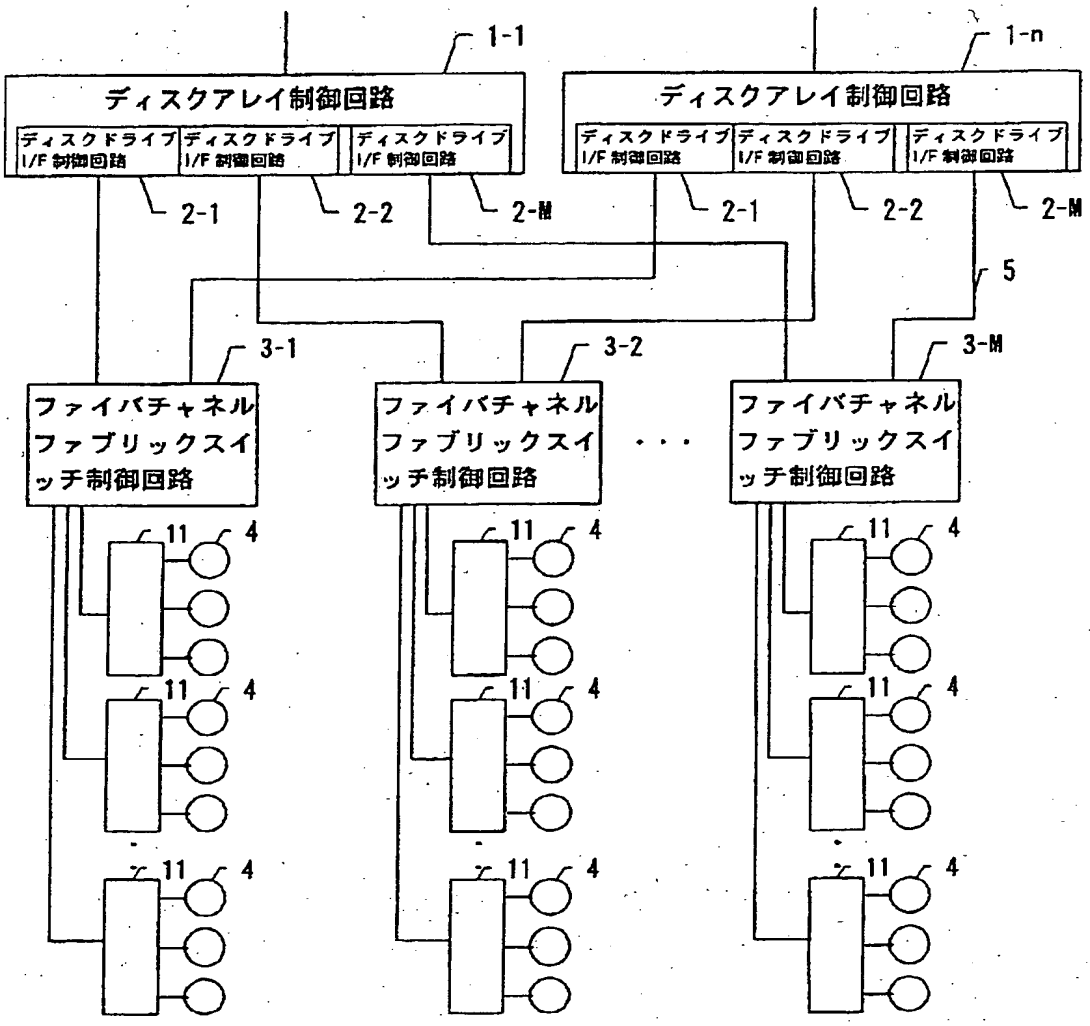
【図４】

図４


1-1
1-n
ディスクアレイ制御回路
ディスクアレイ制御回路
ディスクドライブ I/F 制御回路
ディスクドライブ I/F 制御回路
ディスクドライブ I/F 制御回路
ディスクドライブ I/F 制御回路
ディスクドライブ I/F 制御回路
ディスクドライブ I/F 制御回路
2-1
2-2
2-M
2-1
2-2
2-M
5
3-1
3-2
3-M
10-1
4(1, 1)
4(2, 1)
4(M, 1)
4(1, 2)
4(2, 2)
4(M, 2)
10-L
4(1, L)
4(2, L)
4(M, L)
6

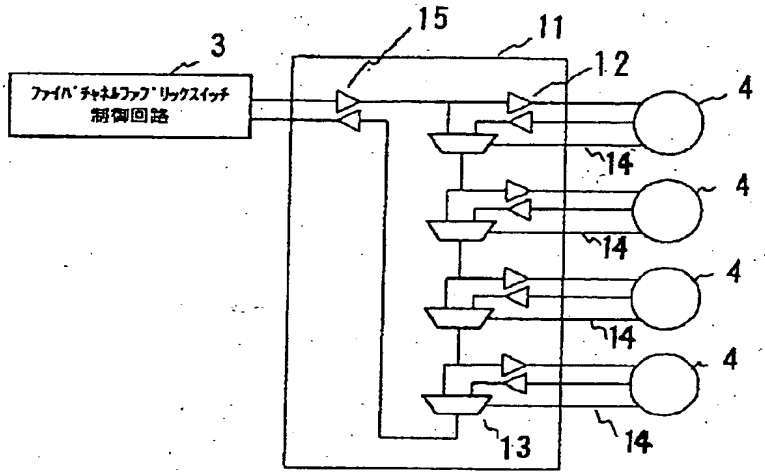
【図５】

図５


1-1
1-n
ディスクアレイ制御回路
ディスクドライブ I/F 制御回路
ディスクドライブ I/F 制御回路
ディスクドライブ I/F 制御回路
2-1
2-2
2-M
5
3-1
3-2
3-M
ファイバチャネル ファブリックスイッチ制御回路
11
4

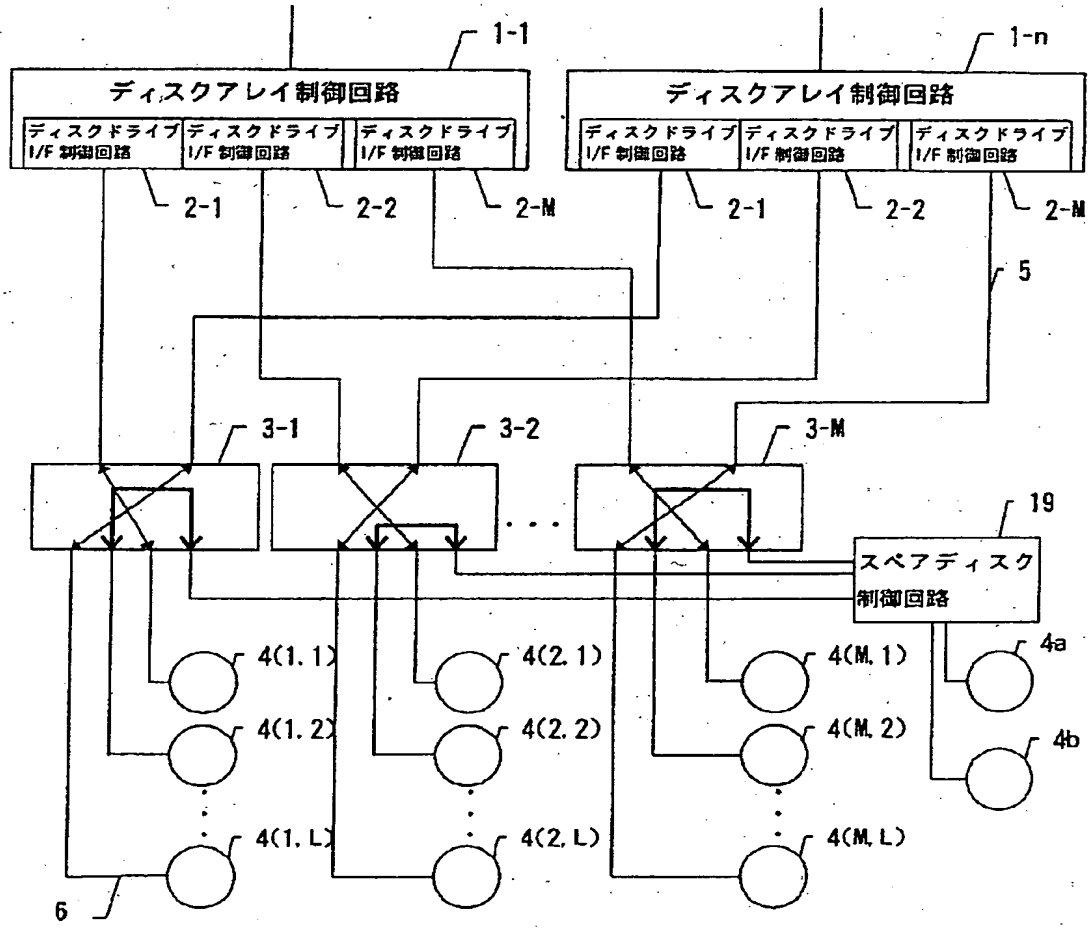
【図６】

図６



【図７】

図７



フロントページの続き

(51) Int.Cl.[7]	識別記号	F I	ﾃｰﾏｺｰﾄﾞ（参考）
Ｇ０６Ｆ　13/10	３４０	Ｇ０６Ｆ　13/10	３４０Ａ

[Patent Document]

1. Japanese Patent Laid Open No. 2003-303055

Disk storage system having disk arrays connected with disk adaptors through switches

Hitachi, Ltd.

Inventor(s):Tanaka, Katsuya ;Fujimoto, Kazuhisa

Application No. 10/212882, Filed 20020807, A1 Published 20031009

Abstract:

A disk storage system has high throughput between a disk adapter of a disk controller and a disk array. The disk adapter of the disk controller is connected to the disk array through switches. Data on a channel between the switch and a RAID group is multiplexed in the switch to be transferred onto a channel between the switch and the disk adapter and data on the channel between the switch and the disk adapter is demultiplexed in the switch to be transferred onto the channel between the switch and the RAID group. A data transfer rate on the channel between the disk adapter and the switch is made higher than that on the channel.

US.Class: 711114 711154

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2003－303055
(P2003－303055A)
(43)公開日　平成15年10月24日(2003.10.24)

(51)Int.Cl.'	識別記号	ＦＩ	テーマコード'(参考)
Ｇ０６Ｆ　3/06	３０１	Ｇ０６Ｆ　3/06	３０１Ｍ　５Ｂ０６５
			３０１Ｂ
	３０２		３０２Ａ
	５４０		５４０

審査請求　未請求　請求項の数９　ＯＬ　（全 14 頁）

(21)出願番号　特願2002－106262(P2002－106262)

(22)出願日　平成14年４月９日(2002.4.9)

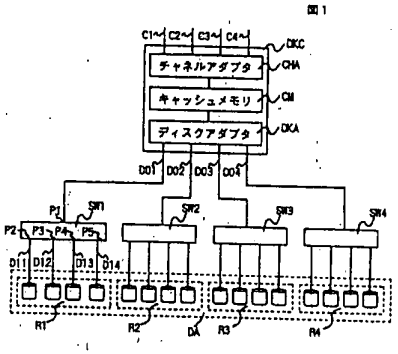
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　田中　勝也
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内
(72)発明者　藤本　和久
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内
(74)代理人　100099298
弁理士　伊藤　修　（外１名）

Ｆターム(参考)　5B065 BA01 CA04 CA07 CA12 CA15
CA30 CC08 CE11 CH01

(54)【発明の名称】　ディスクアダプタとディスクアレイをスイッチを介して接続したディスク装置

(57)【要約】
【課題】　ディスクコントローラのディスクアダプタとディスクアレイ間のスループットが高いディスク装置を提供することにある。
【解決手段】　ディスクコントローラ(DKC)のディスクアダプタ(DKA)とディスクアレイ(DA)をスイッチ(SW1,SW2,SW3,SW4)を介して接続する。スイッチ(SW1)とＲＡＩＤグループ(R1)間のチャネル(D11,D12,D13,D14)上のデータをスイッチ(SW1)において多重化してスイッチ(SW1)とディスクアダプタ(DKA)間のチャネル(D01)に転送し、スイッチ(SW1)とディスクアダプタ(DKA)間のチャネル(D01)上のデータをスイッチ(SW1)において逆多重化してスイッチ(SW1)とＲＡＩＤグループ(R1)間のチャネル(D11,D12,D13,D14)に転送する。ディスクアダプタ(DKA)とスイッチ(SW1)間のチャネル(D01)上のデータ転送速度を、チャネル(D11,D12,D13,D14)のデータ転送速度より高くする。

図１



【特許請求の範囲】

【請求項１】　ディスクコントローラとディスクアレイからなり、前記ディスクコントローラはチャネルアダプタとキャッシュメモリとディスクアダプタを有するディスク装置において、
前記ディスクアダプタと前記ディスクアレイを、バッファメモリを有するスイッチを介して接続し、
前記スイッチは、前記ディスクアダプタが接続されたポートと前記ディスクアレイを構成するディスクドライブが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行うことを特徴とするディスク装置。

【請求項２】　ディスクコントローラと複数のディスクアレイからなり、前記ディスクコントローラはチャネルアダプタとキャッシュメモリとディスクアダプタを有するディスク装置において、
前記ディスクアレイはループ状に接続した複数のディスクドライブからなり、前記ディスクアダプタと前記複数のディスクアレイとをバッファメモリを有するスイッチを介して接続し、
前記ディスクアダプタと前記スイッチ間のチャネル当りデータ転送速度を、前記スイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、
前記スイッチは、前記ディスクアダプタが接続されたポートと前記複数のディスクアレイが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行うことを特徴とするディスク装置。

【請求項３】　ディスクコントローラとディスクアレイからなり、前記ディスクコントローラはチャネルアダプタとキャッシュメモリとディスクアダプタを有するディスク装置において、
前記ディスクアダプタと前記ディスクアレイを、バッファメモリを有するスイッチを介して接続し、
同一のスイッチに接続したディスクドライブの組み合わせでＲＡＩＤグループを構成し、
前記ディスクアダプタと前記スイッチ間のチャネル当りデータ転送速度を、前記スイッチと前記ディスクアレイ間のチャネル当りデータ転送速度より高く設定し、
前記スイッチは、前記ディスクアダプタが接続されたポートと前記ＲＡＩＤグループを構成するディスクドライブが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行うことを特徴とするディスク装置。

【請求項４】　第１のディスクコントローラと第２のディスクコントローラと複数のディスクアレイからなり、
第１のディスクコントローラは第１のチャネルアダプタと第１のキャッシュメモリと第１のディスクアダプタを有し、
第２のディスクコントローラは第２のチャネルアダプタと第２のキャッシュメモリと第２のディスクアダプタを有するディスク装置において、
第１のディスクアダプタと前記複数のディスクアレイとをバッファメモリを有する第１のスイッチを介して接続し、且つ第２のディスクアダプタと前記複数のディスクアレイとをバッファメモリを有する第２のスイッチを介して接続し、さらに第１のスイッチと第２のディスクアダプタを接続し、第２のスイッチと第１のディスクアダプタを接続し、
第１のディスクアダプタと第１のスイッチ間、および第２のディスクアダプタと第１のスイッチ間のチャネル当りデータ転送速度を第１のスイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、
第２のディスクアダプタと第２のスイッチ間、および第１のディスクアダプタと第２のスイッチ間のチャネル当りデータ転送速度を第２のスイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、
第１のスイッチは、第１のディスクアダプタまたは第２のディスクアダプタが接続されたポートと前記複数のディスクアレイが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行い、
第２のスイッチは、第１のディスクアダプタまたは第２のディスクアダプタが接続されたポートと前記複数のディスクアレイが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行うことを特徴とするディスク装置。

【請求項５】　第１のディスクコントローラと第２のディスクコントローラと複数のディスクアレイからなり、
第１のディスクコントローラは第１のチャネルアダプタと第１のキャッシュメモリと第１のディスクアダプタを有し、
第２のディスクコントローラは第２のチャネルアダプタと第２のキャッシュメモリと第２のディスクアダプタを有するディスク装置において、
第１のディスクアダプタと前記複数のディスクアレイとをバッファメモリを有する第１のスイッチを介して接続し、且つ第２のディスクアダプタと前記複数のディスクアレイとをバッファメモリを有する第２のスイッチを介して接続し、さらに第１のスイッチと第２のディスクアダプタを接続し、第２のスイッチと第１のディスクアダプタを接続し、
第１のディスクアダプタと第１のスイッチ間、および第２のディスクアダプタと第１のスイッチ間のチャネル当

りデータ転送速度を第１のスイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、
第２のディスクアダプタと第２のスイッチ間、および第１のディスクアダプタと第２のスイッチ間のチャネル当りデータ転送速度を第２のスイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、
第１のスイッチと第２のスイッチを、第１のディスクアダプタと第２のスイッチ間を接続したチャネルと同等のデータ転送速度を有するチャネルと、第２のディスクアダプタと第１のスイッチ間を接続したチャネルと同等のデータ転送速度を有するチャネルと、を介して接続し、
第１のスイッチは、第１のディスクアダプタまたは第２のディスクアダプタまたは第２のスイッチが接続されたポートと前記複数のディスクアレイが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行い、
第２のスイッチは、第１のディスクアダプタまたは第２のディスクアダプタまたは第１のスイッチが接続されたポートと前記複数のディスクアレイが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行うことを特徴とするディスク装置。
【請求項６】　請求項１乃至請求項５のいずれかの請求項記載のディスク装置において、
前記ディスクアレイからのデータ読み出し時には、前記ディスクアレイから前記スイッチに転送されるデータを前記スイッチにおいて多重化して前記ディスクアダプタに転送し、
前記ディスクアレイへのデータ書き込み時には、前記ディスクアダプタから前記スイッチに転送されるデータを前記スイッチにおいて逆多重化して前記ディスクアレイに転送することを特徴とするディスク装置。
【請求項７】　請求項１乃至請求項５のいずれかの請求項記載のディスク装置において、
ディスグアダプタからディスクアレイへのデータ書き込み時に、前記ディスクアダプタは、前記ポート間の接続の切り替えが周期的に行われるように、送出するフレームに送信先情報を設定し、
ディスクアレイからディスクアダプタへのデータ読み出し時に、前記スイッチは、ラウンドロビン方式により前記ポート間の接続を切り替えることを特徴とするディスク装置。
【請求項８】　請求項７記載のディスク装置において、周期的に切り替えるポート数を、ディスクアダプタとスイッチ間のチャネル当りデータ転送速度の、スイッチとディスクアレイ間のチャネル当りデータ転送速度に対する比、と同程度に設定することを特徴とするディスク装置。
【請求項９】　請求項１乃至請求項５のいずれかの請求項記載のディスク装置において、
前記ディスクアダプタと前記スイッチ間を光ファイバケーブルで接続し、前記スイッチと前記ディスクアレイ間をメタルケーブルで接続することを特徴とするディスク装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、コンピュータシステムにおける２次記憶装置に関し、特に入出力データ転送性能が高いディスク装置に関する。
【０００２】
【従来の技術】現在のコンピュータシステムにおいては、ＣＰＵ（中央処理装置）が必要とするデータは２次記憶装置に保存され、ＣＰＵなどが必要とするときに応じて２次記憶装置に対してデータの書き込みおよび読み出しを行う。この２次記憶装置としては、一般に不揮発な記憶媒体が使用され、代表的なものとして磁気ディスク装置や、光ディスクなどのディスク装置がある。近年高度情報化に伴い、コンピュータシステムにおいて、この種の２次記憶装置の高性能化が要求されている。
【０００３】図９に、従来のディスク装置のブロック図を示す。図９において、ディスク装置はディスクコントローラＤＫＣとディスクアレイＤＡで構成される。ディスクコントローラＤＫＣは、上位側ＣＰＵ（図示せず）とディスク装置を接続するチャネルアダプタＣＨＡと、ディスクアレイＤＡに対して読み書きするデータを一時保存するキャッシュメモリＣＭと、ディスクコントローラＤＫＣとディスクアレイＤＡを接続するディスクアダプタＤＫＡからなる。チャネルアダプタＣＨＡとキャッシュメモリＣＭとディスクアダプタＤＫＡは、バスまたはスイッチで相互接続されている。チャネルアダプタＣＨＡはＣ１、Ｃ２、Ｃ３、Ｃ４の４本のチャネルでＣＰＵと接続している。ディスクアダプタＤＫＡはＤ１、Ｄ２、Ｄ３、Ｄ４の４本のチャネルでディスクアレイと接続している。ここでディスクアレイＤＡはディスクグループＲ１、Ｒ２、Ｒ３、Ｒ４からなる。ディスクアレイＤＡにおいてＲＡＩＤシステムを構築する場合は、Ｒ１、Ｒ２、Ｒ３、Ｒ４がそれぞれＲＡＩＤグループを構成する。
【０００４】チャネルＣ１、Ｃ２、Ｃ３、Ｃ４から入力された書き込みデータは、キャッシュメモリＣＭに該データを書き込むと同時に、該データをブロックサイズ単位に分割し、チャネルＤ１、Ｄ２、Ｄ３、Ｄ４の内３チャネルにはブロック単位に分割されたデータを、残りの１チャネルは前記分割データから計算したパリティを、ディスクアダプタＤＫＡからディスクアレイＤＡへ送信する。データ読み出し時は、先ずキャッシュメモリＣＭ内に該当データの有無を調べる。有る場合は、キャッシュメモリＣＭからチャネルアダプタＣＨＡを介してキャッシュメモリ内読み出しデータをＣＰＵへ送信する。キャッシュメモリＣＭ内に無い場合にディスクアダプタＤＫＡは、Ｄ１、Ｄ２、Ｄ３、Ｄ４を介してディスクアレイＤＡからブロック単位に分割されたデータを読み出し、チャネルアダプタＣＨＡを介して読み出しデータをＣＰＵへ送信する。　この種の従来技術を第１の従来技術と呼ぶ。第１の従来技術として関連するディスク装置は、例えば、日経ＢＰ社刊の「日経コンピュータ別冊メインフレーム’９８」（１９９８年）第１４４頁から第１５３頁に記載されているディスク装置がある。
【０００５】ディスクアダプタとディスクアレイを、スイッチを介して接続したディスク装置が、特開平５－１７３７２２号の「マルチチャンネルデータおよびパリティの交換デバイス」に開示されている。以下、該公報に記載の従来の技術を第２の従来技術と呼ぶ。第２の従来の技術によれば、ディスクアレイに関連したパス本数とディスクアダプタに関連したパス本数とを独立に設定できる。ディスクアダプタとディスクアレイを、バッファ制御ブロックを介して接続したディスク装置が、特開平６－１９６２７号の「回転形記憶装置」に開示されている。以下、該公報に記載の従来の技術を第３の従来技術と呼ぶ。第３の従来技術によれば、ディスクアダプタとディスクアレイ間のデータ転送速度を任意に設定でき、ディスクの回転待ちの影響を低減できる。
【０００６】
【発明が解決しようとする課題】ネットワーク技術の進歩に伴い、１チャネル当りのデータ転送速度は年々増加している。例えばディスク装置に使われるファイバチャネルでは、現状でチャネル当りのデータ転送速度が１Ｇｂｐｓから２Ｇｂｐｓであるが、近い将来４Ｇｂｐｓから１０Ｇｂｐｓへ高速化されることが予定されている。ＣＰＵとチャネルアダプタ間（以下フロントエンドと呼ぶ）のスループットはこの高速化に従うことが予想される。ところが、ディスクアダプタとディスクアレイ間（以下バックエンドと呼ぶ）のスループットは以下の理由により、フロントエンドほど高速化されないと予想される。第１の理由は、ディスクドライブは機械部品を含むので、電子、光素子のみ高速化を行えば良いフロントエンドに比べ高速化が難しいこと、である。第２の理由は、たとえディスクドライブが高速化したとしても、全てのディスクドライブ毎に高速インターフェイスを搭載するのは、多数のディスクドライブを有するディスク装置の高コスト化を招くことである。
【０００７】第１の従来技術では、チャネルアダプタのチャネル当りのデータ転送速度を向上させても、フロントエンドとバックエンドのスループット乖離により、ディスク装置の性能が向上しないという問題があった。また、バックエンドのスループット向上のために低速ポートを多数、ディスクアダプタに設けることも考えられるが、ディスクアダプタのポート数増加は制御を複雑とする。第２の従来技術では、ディスクアダプタとディスクアレイとの間にスイッチを適用することによりディスク増設ポート数を増加させることができるが、チャネル当りのデータ転送速度はディスクアレイのデータ転送速度に制限されるので、ディスクアダプタとディスクアレイ間のスループットが性能ネックになるという問題があった。第３の従来技術は、ディスクの回転待ち時間の影響を低減できる技術であり、フロントエンドとバックエンドのスループット乖離は低減できないという問題があった。
【０００８】本発明の目的は、ディスクアダプタとディスクアレイ間のスループットが高いディスク装置を提供することにある。本発明の他の目的は、ディスクアダプタとディスクアレイ間のスループットが高く、且つディスクドライブ接続台数が多いディスク装置を提供することにある。本発明のさらに他の目的は、信頼性が高いディスクアレイを有するディスク装置を提供することである。本発明のさらに他の目的は、信頼性が高いディスクアダプタとディスクアレイ間ネットワークを有するディスク装置を提供することにある。本発明のさらに他の目的は、信頼性およびスループットが高いディスクアダプタとディスクアレイ間ネットワークを有するディスク装置を提供することにある。本発明のさらに他の目的は、ディスクからの読み出しおよびディスクへの書きこみを高スループット化できるディスク装置を提供することにある。本発明のさらに他の目的は、高スループットを維持できるディスク装置を提供することである。本発明のさらに他の目的は、高スループットで低コストなディスク装置を提供することである。
【０００９】
【課題を解決するための手段】上記目的を達成するため、本発明は、ディスクコントローラとディスクアレイからなり、ディスクコントローラはチャネルアダプタとキャッシュメモリとディスクアダプタを有するディスク装置であり、ディスクアダプタとディスクアレイを、バッファメモリを有するスイッチを介して接続し、ディスクアダプタとスイッチ間のチャネル当りデータ転送速度を、スイッチとディスクアレイ間のチャネル当りデータ転送速度より高く設定し、スイッチは、ディスクアダプタが接続されたポートとディスクアレイを構成するディスクドライブが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行っている。また、前記ディスクアレイはループ状に接続した複数のディスクドライブからなり、前記ディスクアダプタと前記複数のディスクアレイとをバッファメモリを有するスイッチを介して接続し、ディスクアダプタとスイッチ間のチャネル当りデータ転送速度を、スイッチと複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、

スイッチは、ディスクアダプタが接続されたポートと複数のディスクアレイが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行っている。また、前記ディスクアダプタと前記ディスクアレイを、バッファメモリを有するスイッチを介して接続し、同一のスイッチに接続したディスクドライブの組み合わせでＲＡＩＤグループを構成し、ディスクアダプタとスイッチ間のチャネル当りデータ転送速度を、スイッチとディスクアレイ間のチャネル当りデータ転送速度より高く設定し、スイッチは、ディスクアダプタが接続されたポートとＲＡＩＤグループを構成するディスクドライブが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行っている。また、第１のディスクコントローラと第２のディスクコントローラと複数のディスクアレイからなり、第１のディスクコントローラは第１のチャネルアダプタと第１のキャッシュメモリと第１のディスクアダプタを有し、第２のディスクコントローラは第２のチャネルアダプタと第２のキャッシュメモリと第２のディスクアダプタを有するディスク装置であり、第１のディスクアダプタと前記複数のディスクアレイとをバッファメモリを有する第１のスイッチを介して接続し、且つ第２のディスクアダプタと前記複数のディスクアレイとをバッファメモリを有する第２のスイッチを介して接続し、さらに第１のスイッチと第２のディスクアダプタを接続し、第２のスイッチと第１のディスクアダプタを接続し、第２のディスクアダプタと第２のスイッチ間、および第１のディスクアダプタと第２のスイッチ間のチャネル当りデータ転送速度を第２のスイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、第１のスイッチは、第１のディスクアダプタまたは第２のディスクアダプタが接続されたポートと前記複数のディスクアレイが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行い、第２のスイッチは、第１のディスクアダプタまたは第２のディスクアダプタが接続されたポートと前記複数のディスクアレイが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行っている。また、さらに、上記第１のスイッチと第２のスイッチを、上記第１のディスクアダプタと第２のスイッチ間を接続したチャネルと同等のデータ転送速度を有するチャネルと、第２のディスクアダプタと第１のスイッチ間を接続したチャネルと同等のデータ転送速度を有するチャネルと、を介して接続している。また、前記ディスクアレイからのデータ読み出し時には、前記ディスクアレイから前記スイッチに転送されるデータを前記スイッチにおいて多重化して前記ディスクアダプタに転送し、前記ディスクアレイへのデータ書き込み時には、前記ディスクアダプタから前記スイッチに転送されるデータを前記スイッチにおいて逆多重化して前記ディスクアレイに転送するようにしている。また、前記ディスクアダプタから前記ディスクアレイへのデータ書き込み時に、前記ディスクアダプタは、前記ポート間の接続の切り替えが行われるように、送出するフレームに送信先情報を設定し、前記ディスクアレイから前記ディスクアダプタへのデータ読み出し時に、前記スイッチは、ラウンドロビン方式により前記ポート間の接続を切り替えるようにしている。また、さらに、切り替えるポート数を、ディスクアダプタとスイッチ間のチャネル当りデータ転送速度の、スイッチとディスクアレイ間のチャネル当りデータ転送速度に対する比、と同程度に設定している。また、前記ディスクアダプタと前記スイッチ間を光ファイバケーブルで接続し、前記スイッチと前記ディスクアレイ間をメタルケーブルで接続するようにしている。

【００１０】

【発明の実施の形態】以下、図面を参照して本発明の実施の形態を詳細に説明する。図１に本発明の、第１の実施の形態であるディスク装置の構成を示す。本実施の形態のディスク装置は、ディスクコントローラＤＫＣとディスクアレイＤＡからなる。ディスクコントローラＤＫＣは、チャネルアダプタＣＨＡと、キャッシュメモリＣＭと、ディスクアダプタＤＫＡからなる。チャネルアダプタＣＨＡは、上位ＣＰＵ（図示せず）とディスクコントローラＤＫＣとがデータを送受信する際の制御を行う。Ｃ１、Ｃ２、Ｃ３およびＣ４は、チャネルアダプタＣＨＡがＣＰＵと通信するチャネルである。キャッシュメモリＣＭは、本実施の形態のディスク装置が入出力するデータを一時保存するメモリである。ディスクアダプタＤＫＡは、ディスクコントローラＤＫＣとディスクアレイＤＡとがデータを送受信する際の制御を行う。ディスクアダプタＤＫＡは、チャネルＤ０１、Ｄ０２、Ｄ０３、Ｄ０４を介して、ディスクアレイＤＡと接続する。ディスクアダプタＤＫＡとディスクアレイＤＡは、チャネルＤ０１、Ｄ０２、Ｄ０３、Ｄ０４上で全二重通信が可能である。

【００１１】ここで、本実施の形態のディスク装置は、ディスクアダプタＤＫＡとディスクアレイＤＡを、スイッチＳＷ１、ＳＷ２、ＳＷ３、ＳＷ４を介して接続している点に特徴がある。ディスクアレイＤＡは、ディスクグループＲ１、Ｒ２、Ｒ３、Ｒ４からなる。ディスクグループＲ１は、スイッチＳＷ１介してディスクアダプタＤＫＡと接続する。同様に、ディスクグループＲ２はスイッチＳＷ２を介して、ディスクグループＲ３はスイッチＳＷ３を介して、ディスクグループＲ４はスイッチＳＷ４介して、それぞれディスクアダプタＤＫＡと接続する。

【００１２】本実施の形態のディスク装置においてＲＡＩＤシステムを構築する場合は、ディスクグループＲ１、Ｒ２、Ｒ３、Ｒ４を、それぞれＲＡＩＤグループとする。本実施の形態では、４個のディスクドライブでＲＡＩＤグループを構成しているが、ＲＡＩＤグループを構成するドライブ数を４個に限るものではない。各ディスクグループへのデータ読み出しまたは書き込み時のデータの流れを、ディスクグループＲ１を例にして述べる。ここでＲ１はＲＡＩＤレベル５のＲＡＩＤグループである。チャネルＣ１、Ｃ２、Ｃ３、Ｃ４からディスクグループＲ１へ書き込むためにＣＰＵから送信されたデータは、ディスクアダプタＤＫＡにおいてブロック単位に分割されると同時に、該ブロック単位に分割されたデータからパリティが生成される。該ブロック単位に分割されたデータと、生成されたパリティは、チャネルＤ０１を通りスイッチＳＷ１へ入力される。スイッチＳＷ１は、ＲＡＩＤ制御に伴い、該ブロック単位に分割されたデータと、生成されたパリティとをルーティングし、チャネルＤ１１、Ｄ１２、Ｄ１３、Ｄ１４へ分配する。データ読み出し時は、ディスクアダプタＤＫＡは、Ｄ１１、Ｄ１２、Ｄ１３、Ｄ１４を介してディスクグループＲ１からブロック単位に分割されたデータを読み出し、スイッチＳＷ１でシリアル化して、チャネルＤ０１を介して読み出しデータを受信する。

【００１３】図９に示した従来のディスク装置では、ディスクアダプタＤＫＡに接続したチャネルＤ１、Ｄ２、Ｄ３、Ｄ４上で、既にディスクアレイへの書き込みデータおよびパリティが別々のチャネルに分配されていた。それに対し、本実施の形態のディスク装置においては、スイッチＳＷ１通過後に別々のチャネルに分配される点が従来と異なる。

【００１４】次に、本実施の形態のディスク装置の特徴であるスイッチの動作を、スイッチＳＷ１を例にとり説明する。ＳＷ２～ＳＷ４の動作もＳＷ１の動作と同様である。図１に示すように、スイッチＳＷ１は入出力ポートＰ１、Ｐ２、Ｐ３、Ｐ４、Ｐ５を有する。ポートＰ１、Ｐ２、Ｐ３、Ｐ４、Ｐ５は、全二重通信可能な入出力ポートであり、ポート毎にバッファメモリを有している。スイッチＳＷ１の内部構成を図２と図３に示す。簡単のため、データの進行方向によりスイッチ動作を分けて説明する。また、チャネルＤ０１、Ｄ１１、Ｄ１２、Ｄ１３、Ｄ１４上を流れるデータは、フレーム単位で送受信され、かつデータは８Ｂ１０Ｂ変換で符号化されている。

【００１５】図２は、ポートＰ１からブロック内のフレームを入力し、ポートＰ２、Ｐ３、Ｐ４、Ｐ５から出力する場合を示す。これはディスクアレイへの書き込み時のスイッチ動作に相当する。スイッチＳＷ１は図２に示すように、クロスバスイッチＸＳＷと、スイッチコントローラＣＴＬからなる。クロスバスイッチＸＳＷは５×５のクロスバスイッチであり、入力ポートｉｎ１、ｉｎ２、ｉｎ３、ｉｎ４、ｉｎ５と、出力ポートｏｕｔ１、ｏｕｔ２、ｏｕｔ３、ｏｕｔ４、ｏｕｔ５を有する。ポートＰ１から入力したフレームは、シリアルパラレル変換装置ＳＰ１と、バッファメモリＢＭ１と、８Ｂ１０Ｂ変換デコーダＤＥＣ１を経由し、スイッチコントローラＣＴＬと入力ポートｉｎ１へ入力される。スイッチコントローラＣＴＬにおいて、入力フレームのヘッダ部分に書かれた送信先アドレスを解読し、クロスバスイッチＸＳＷを切り換える。例として、ポートＰ２が出力先として選ばれた場合は、入力したフレームは出力ポートｏｕｔ２と、８Ｂ１０Ｂ変換エンコーダＥＮＣ２と、バッファメモリＢＭ２と、パラレルシリアル変換装置ＰＳ２を経由して、ポートＰ２から出力される。ここで、バッファメモリＢＭ１、ＢＭ２はＦＩＦＯ（Ｆｉｒｓｔ－ｉｎ　Ｆｉｒｓｔ－ｏｕｔ）メモリである。

【００１６】シリアルパラレル変換装置ＳＰ１は、８Ｂ１０Ｂ符号化されたシリアルデータを１０ｂｉｔ幅のパラレルデータに変換し、ポートＰ１におけるデータ転送速度の１／１０の速度に同期してバッファメモリＢＭ１に書き込む。８Ｂ１０Ｂデコーダ ＤＥＣ１は、クロスバスイッチＸＳＷの動作速度に同期して、１０ｂｉｔパラレルデータをバッファメモリＢＭ１から読み出し、８Ｂ１０Ｂ復号化して、８ｂｉｔパラレルデータに変換する。８Ｂ１０ＢエンコーダＥＮＣ２は、クロスバスイッチＸＳＷでスイッチされた８ｂｉｔパラレルデータを再び８Ｂ１０Ｂ符号化し、１０ｂｉｔパラレルデータに変換後、クロスバスイッチＸＳＷの動作速度に同期してバッファメモリＢＭ２に書き込む。パラレルシリアル変換装置ＰＳ２は、ポートＰ２におけるデータ転送速度の１／１０の速度に同期して、１０ｂｉｔパラレルデータをバッファメモリＢＭ２から読み出し、シリアル化して、ポートＰ２から出力する。以上によりスイッチＳＷ１は、ポートＰ１におけるデータ転送速度からポートＰ２におけるデータ転送速度へ速度変換する。

【００１７】図４は、ポートＰ１へ入力するフレームと、ポートＰ２、Ｐ３、Ｐ４、Ｐ５から出力されるフレームを示した図である。波形の凸はフレームが存在する時間、凹はフレームが存在していない時間を示している。フレームは伝送するデータ容量に従ってそのフレーム長が変化するが、ここではディスクアレイへのシーケンシャルアクセスが行われており、フレーム長が一定である。図４では、入力ポートＰ１でのデータ転送速度が出力ポートＰ２、Ｐ３、Ｐ４、Ｐ５におけるデータ転送速度のｍ倍あるとする。従って、ポートＰ１におけるフレームＦｂ２の時間Ｔ１は、ポートＰ２からの出力時にＴ３へ伸びている。ここでＴ３＝ｍ×Ｔ１である。

【００１８】入力のデータ転送速度が速く、且つ出力のデータ転送速度が遅い場合は、スイッチを周期的に切り換えないと出力ポートのバッファメモリが溢れ、スループットが低下する。フレームがスループットの低下無く

スイッチを通過するには、図４のように周期的に出力ポートを切り換える必要がある。スイッチ切替えポート数をｎとすると、スイッチ切替え周期Ｔ２≒ｎ×Ｔ１である（フレームの無い時間は無視した）。Ｔ２≧Ｔ３ならば、フレームの衝突無く、スループットの低下は起こらない。Ｔ２≧Ｔ３はｎ≧ｍと同じある。つまり、ディスクアレイへのデータ書き込み時に、スイッチにおいてスループット低下を起こさないための条件は、周期的に切り替えるスイッチポート数ｎを、ディスクアダプタとスイッチ間のチャネル当りデータ転送速度の、スイッチとディスクアレイ間のチャネル当りデータ転送速度に対する比ｍ、以上に設定することである。この条件が保たれれば、スイッチＳＷ１は、ポートＰ１から入力したデータをバッファメモリにおいて速度変換し、フレーム単位で周期的に切り替えることにより逆多重化し、ポートＰ２、Ｐ３、Ｐ４、Ｐ５へ分配して出力する。スイッチを周期的に切り換える方法の一つは、スイッチに接続したディスクグループをＲＡＩＤグループとすることである。ＲＡＩＤのストライピング制御に従えば、スイッチは周期的に切り替わる。

【００１９】図３は、ポートＰ２、Ｐ３、Ｐ４、Ｐ５からフレームを入力し、ポートＰ１から出力する場合を示す。これはディスクアレイからの読み出し時のスイッチ動作に相当する。例えば、ポートＰ２から入力したフレームは、シリアルパラレル変換装置ＳＰ２と、バッファメモリＢＭ２と、８Ｂ１０Ｂ変換デコーダＤＥＣ２を経由し、スイッチコントローラＣＴＬと入力ポートｉｎ２へ入力される。スイッチコントローラＣＴＬにおいて、入力フレームのヘッダ部分に書かれた送信先アドレスを解読し、クロスバスイッチＸＳＷを切り換える。図３の場合は、ラウンドロビン方式によりクロスバススイッチＸＳＷを切り替えて、順番にポートＰ２、Ｐ３、Ｐ４、Ｐ５から入力されるデータは全てポートＰ１へ出力する。すなわち、読み出し時は、複数の入力ポート（Ｐ２、Ｐ３、Ｐ４、Ｐ５）に同時にフレームが届く。これら複数の入力フレームは同期して入力ポートに届く必要はない。スイッチは、総当り的に入出力ポート間接続を切り替えることにより、これら複数の入力フレームを１フレームずつ出力ポート（Ｐ１）へ転送する。このように、スイッチを総当り的に切り替える方式を、ラウンドロビン(Round Robin)方式と呼ぶ。ラウンドロビン方式により、結果的にスイッチは周期的に切り替わることになる。なお、読み出し時においても、スイッチはフレーム内送信先情報に従って切り替わることに違いはない。フレームは出力ポートｏｕｔ１と、８Ｂ１０Ｂ変換エンコーダＥＮＣ１と、バッファメモリＢＭ１と、パラレルシリアル変換装置ＰＳ１を経由して、ポートＰ１から出力される。

【００２０】シリアルパラレル変換装置ＳＰ２は、８Ｂ１０Ｂ符号化されたシリアルデータを１０ｂｉｔ幅のパラレルデータに変換し、ポートＰ２におけるデータ転送速度の１／１０の速度に同期してバッファメモリＢＭ２に書き込む。８Ｂ１０Ｂデコーダ ＤＥＣ２は、クロスバスイッチＸＳＷの動作速度に同期して、１０ｂｉｔパラレルデータをバッファメモリＢＭ２から読み出し、８Ｂ１０Ｂ復号化して、８ｂｉｔパラレルデータに変換する。８Ｂ１０Ｂエンコーダ ＥＮＣ１は、クロスバスイッチＸＳＷでスイッチされた８ｂｉｔパラレルデータを再び８Ｂ１０Ｂ符号化し、１０ｂｉｔパラレルデータに変換後、クロスバスイッチＸＳＷの動作速度に同期してバッファメモリＢＭ１に書き込む。パラレルシリアル変換装置ＰＳ１は、ポートＰ１におけるデータ転送速度の１／１０の速度に同期して、１０ｂｉｔパラレルデータをバッファメモリＢＭ１から読み出し、シリアル化して、ポートＰ１から出力する。以上によりスイッチＳＷ１は、ポートＰ２におけるデータ転送速度からポートＰ１におけるデータ転送速度へ速度変換する。

【００２１】図５は、ポートＰ２、Ｐ３、Ｐ４、Ｐ５へ入力するフレームと、ポートＰ１から出力されるフレームを示した図である。波形の凸はフレームが存在する時間、凹はフレームが存在していない時間を示している。フレームは伝送するデータ容量に従ってそのフレーム長が変化するが、ここではディスクアレイへのシーケンシャルアクセスが行われており、フレーム長が一定である。図５では、入力ポートＰ１でのデータ転送速度が出力ポートＰ２、Ｐ３、Ｐ４、Ｐ５におけるデータ転送速度のｍ倍あるとする。従って、ポートＰ５におけるフレームＦｅ５の時間Ｔ４は、ポートＰ１からの出力時にＴ５へ縮んでいる。ここでＴ４＝ｍ×Ｔ５である。フレームＦｅ２、Ｆｅ３、Ｆｅ４、Ｆｅ５をポートＰ１から出力するのにかかる時間をＴ６とする。スイッチ切り替えポート数をｎとすると、Ｔ６≒ｎ×Ｔ５である（フレームの無い時間は無視した）。スイッチにおいて輻輳によるスループット低下を防止するためには、Ｔ６≦Ｔ４とする必要がある。Ｔ６≦Ｔ４はｎ≦ｍと同じある。

【００２２】つまり、ディスクアレイからのデータ読み出し時に、スイッチにおいてスループット低下を起こさないための条件は、周期的に切り替えるスイッチポート数ｎを、ディスクアダプタとスイッチ間のチャネル当りデータ転送速度の、スイッチとディスクアレイ間のチャネル当りデータ転送速度に対する比ｍ、以下に設定すること、である。この条件が保たれれば、スイッチＳＷ１は、ポートＰ２、Ｐ３、Ｐ４、Ｐ５から入力したデータをバッファメモリにおいて速度変換し、フレーム単位で周期的に切り替えることにより多重化し、ポートＰ１へ出力する。よって、ディスクアレイへの書き込みおよびディスクアレイからの読み出しを高スループット化するためには、ｎ≒ｍ、つまり、周期的に切り替えるポート数を、ディスクアダプタとスイッチ間のチャネル当りデータ転送速度の、スイッチとディスクアレイ間のチャネル当りデータ転送速度に対する比、と同程度に設定すればよいことが分かる。

【００２３】例えば、ディスクアダプタとスイッチ間の４Ｇｂｐｓのチャネル１本で接続し、スイッチとディスクアレイ間を１Ｇｂｐｓのチャネル４本で接続する。また、ディスクアダプタとスイッチ間の１０Ｇｂｐｓのチャネル１本で接続し、スイッチとディスクアレイ間を２Ｇｂｐｓのチャネル４本で接続する。この場合、スイッチ入出力ポート間でスループットのバランスが取れないので、実効的なスループットは２Ｇｂｐｓ×４＝８Ｇｂｐｓとなる。

【００２４】以上より、スイッチＳＷ１において速度変換と多重化、逆多重化が行われるので、チャネルＤ１１、Ｄ１２、Ｄ１３、Ｄ１４上のデータ転送速度が低速でも、チャネルＤ０１、Ｄ０２、Ｄ０３、Ｄ０４でのデータ転送速度は高速にできる。つまり、ディスクアダプタＤＫＡとディスクアレイＤＡ間のスループットを向上できる。本実施の形態のディスク装置におけるデータ転送方式としては、ファイバチャネルやインフィニバンドが使用できる。

【００２５】図６は、第１の実施の形態のディスク装置において、ディスクドライブの増設方法を示した図である。図６では図１に対して、ディスクグループＲ５とＲ６が増設されている。ディスクドライブを増設するため、スイッチＳＷ１とＳＷ２としてポート数の多いスイッチを使用している。ディスクドライブを増設すると、スイッチのディスクアレイ側のスループットが増加し、ディスクアダプタ側とのスループットバランスが崩れるので、スイッチの速度変換機能が有効に働かなくなる可能性がある。そこでスイッチＳＷ１では、ディスクアダプタＤＫＡとの間に、新規チャネルＤ０５を増設している。また、スイッチＳＷ２の場合は新規チャネルを増設せず、チャネルＤ０２の信号伝送速度を増加させることで、ディスクアダプタ側とディスクアレイ側のスループットバランスを取っている。例えばスイッチＳＷ１では、スイッチとディスクアレイ間を１Ｇｂｐｓのチャネル８本で接続し、ディスクアダプタとスイッチ間を４Ｇｂｐｓのチャネル２本で接続する。スイッチＳＷ２では、スイッチとディスクアレイ間を１Ｇｂｐｓのチャネル８本で接続し、ディスクアダプタとスイッチ間を１０Ｇｂｐｓのチャネル１本で接続する。このように、本実施の形態のディスク装置は、スイッチのポート数に応じて、ディスクドライブを増設可能である。このディスクドライブ増設方法は、１ポート当たりに接続できるドライブ数が少ないＡＴＡ（ＡＴ Ａｔｔａｃｈｍｅｎｔ）方式ディスクドライブを増設するのに適用できる。

【００２６】図７に本発明の、第２の実施の形態であるディスク装置の構成を示す。本実施の形態のディスク装置は、第１の実施の形態のディスク装置に対して、ディスクアレイ部分の構成方法が異なる。本実施の形態のディスク装置は、ディスクコントローラＤＫＣと、４個のディスクアレイＤＡ１、ＤＡ２、ＤＡ３、ＤＡ４からなる。ディスクコントローラＤＫＣは、チャネルアダプタＣＨＡ、キャッシュメモリＣＭ、ディスクアダプタＤＫＡからなる。ディスクアレイＤＡ１とディスクアダプタＤＫＡは、チャネルＤ０１とスイッチＳＷ１を介して接続する。同様に、ディスクアレイＤＡ２はチャネルＤ０２とスイッチＳＷ２を介して、ディスクアレイＤＡ３はチャネルＤ０３とスイッチＳＷ３を介して、ディスクアレイＤＡ４はチャネルＤ０４とスイッチＳＷ４を介して、それぞれディスクアダプタＤＫＡと接続する。スイッチＳＷ１、ＳＷ２、ＳＷ３とＳＷ４は、第１の実施の形態と同様に速度変換と多重化、逆多重化を行うスイッチとして機能する。本実施の形態におけるディスクアダプタＤＫＡと、スイッチＳＷ１、ＳＷ２、ＳＷ３、ＳＷ４と、ディスクアレイＤＡ１、ＤＡ２、ＤＡ３、ＤＡ４との間のデータ転送方式は、ファイバチャネルを使用している。スイッチＳＷ１、ＳＷ２、ＳＷ３、ＳＷ４はファイバチャネルスイッチである。

【００２７】本実施の形態におけるディスクアレイの構成を、ディスクアレイＤＡ１を例に述べる。ディスクアレイＤＡ１、ＤＡ２、ＤＡ３、ＤＡ４は、同様のドライブ構成である。ディスクアレイＤＡ１は、チャネルＤ１１上に接続した４個のディスクからなるディスクアレイと、Ｄ１２上に接続した４個のディスクからなるディスクアレイと、Ｄ１３上に接続した４個のディスクからなるディスクアレイと、Ｄ１４上に接続した４個のディスクからなるディスクアレイ、からなる。チャネルＤ１１を例にとると、ディスクドライブＤＫ１、ＤＫ２、ＤＫ３、ＤＫ４が、チャネルＤ１１上に接続されている。このように、多数のドライブを一つのチャネル上に接続してディスクドライブにアクセスする方法としては、ファイバチャネルアービトレイテッドループ（以下ＦＣ－ＡＬと呼ぶ）がある。

【００２８】図１０に、ＦＣ－ＡＬの接続形態をディスクドライブＤＫ１、ＤＫ２、ＤＫ３、ＤＫ４の接続形態を例として示す。各ディスクドライブの入出力ポートおよびスイッチＳＷ１の入出力ポートは、送信機Ｔｘと受信機Ｒｘを有する。ＦＣ－ＡＬの接続形態は、例えば図１０に示すように、各ドライブの入出力ポートおよびスイッチの入出力ポートをループ状に接続するトポロジである。各ドライブの入出力ポートはファイバチャネルのＮＬ（Ｎｏｄｅ Ｌｏｏｐ）ポートとして機能する。ＮＬポートとは、ループ動作をする装置（ここではディスクドライブ）のポートである。スイッチＳＷ１のディスクアレイＤＡ１接続側入出力ポートは、ファイバチャネルのＦＬ（Ｆａｂｒｉｃ Ｌｏｏｐ）ポートとして機能する。ＦＬポートとは、ＦＣ－ＡＬを接続可能なスイッチのポートである。ＦＬポートを有するループは、ファイバチャネルのパブリックループとして機能するので、

チャネルＤ１１が形成するＦＣ－ＡＬはパブリックループとなる。パブリックループとは、ループ上のディスクドライブが、スイッチを介してループ外のポートと通信可能なループである。よって、ディスクドライブＤＫ１、ＤＫ２、ＤＫ３、ＤＫ４は、スイッチＳＷ１およびチャネルＤ０１を介してディスクアダプタＤＫＡと通信可能である。以上、チャネルＤ１１の接続形態を例に説明したが、チャネルＤ１２、Ｄ１３、Ｄ１４でも同様である。本実施の形態のディスク装置においてＲＡＩＤシステムを構築する場合は、ディスクグループＲ１、Ｒ２、Ｒ３、Ｒ４を、それぞれＲＡＩＤグループとする。本実施の形態では、４個のディスクドライブでＲＡＩＤグループを構成しているが、ＲＡＩＤグループを構成するドライブ数を４個に限るものではない。

【００２９】本実施の形態においては、チャネルＤ１１、Ｄ１２、Ｄ１３、Ｄ１４において、それぞれＦＣ－ＡＬを用いてディスクドライブを接続している。ＦＣ－ＡＬの仕様から、チャネルＤ１１、Ｄ１２、Ｄ１３、Ｄ１４上には、それぞれ最大１２６台までのディスクドライブが接続可能である。また、チャネルＤ０１、Ｄ０２、Ｄ０３、Ｄ０４の媒体として光ファイバケーブルを、チャネルＤ１１、Ｄ１２、Ｄ１３、Ｄ１４の媒体としてメタルケーブルを用いる。

【００３０】以上説明したように、本実施の形態のディスク装置においては、ディスクドライブをＦＣ－ＡＬで接続しているので、スイッチのポート当りに接続できるドライブ台数が増加できる。つまり、ディスク装置当りの記憶容量を増加させる効果がある。また、ディスクドライブをメタルケーブルで接続することにより、ディスクドライブ毎に高価な光インターフェイスを装備する必要がなくなるので、ディスクドライブのコストを下げる効果がある。

【００３１】図８に本発明の、第３の実施の形態であるディスク装置の構成を示す。本実施の形態のディスク装置は、ディスクコントローラとスイッチを二重化した点に特徴がある。本実施の形態において、ディスクアダプタＤＫＡ１、ＤＫＡ２と、スイッチＳＷ１、ＳＷ２と、ディスクアレイＤＡ１との間のデータ転送方式は、ファイバチャネルを使用している。本実施の形態のディスク装置は、ディスクコントローラＤＫＣ１、ＤＫＣ２と、スイッチＳＷ１、ＳＷ２と、ディスクアレイＤＡ１からなる。スイッチＳＷ１とＳＷ２は、第１の実施の形態と同様に速度変換と多重化、逆多重化を行うスイッチとして機能する。ディスクコントローラＤＫＣ１は、チャネルアダプタＣＨＡ１と、キャッシュメモリＣＭ１と、ディスクアダプタＤＫＡ１からなる。ディスクコントローラＤＫＣ２は、チャネルアダプタＣＨＡ２と、キャッシュメモリＣＭ２と、ディスクアダプタＤＫＡ２からなる。ディスクアダプタＤＫＡ１とスイッチＳＷ１をチャネルＤ１ａで接続し、ディスクアダプタＤＫＡ２とスイッチＳＷ２をチャネルＤ２ａで接続し、ディスクアダプタＤＫＡ１とスイッチＳＷ２をチャネルＤ１ｂで接続し、ディスクアダプタＤＫＡ２とスイッチＳＷ１をチャネルＤ２ｂで接続する。

【００３２】ディスクアレイＤＡ１を構成するディスクドライブは、入出力ポートを２個有する。例えば、ディスクドライブＤＫ１、ＤＫ２、ＤＫ３、ＤＫ４は、チャネルＤ１１およびＤ２１の両チャネルと接続する。ディスクアレイＤＡ１は、チャネルＤ１１とＤ２１に接続した４個のディスクからなるディスクアレイと、Ｄ１２とＤ２２に接続した４個のディスクからなるディスクアレイと、Ｄ１３とＤ２３に接続した４個のディスクからなるディスクアレイと、Ｄ１４とＤ２４に接続した４個のディスクからなるディスクアレイ、からなる。チャネルＤ１１、Ｄ１２、Ｄ１３、Ｄ１４、Ｄ２１、Ｄ２２、Ｄ２３、Ｄ２４は、ＦＣ－ＡＬでディスクドライブを接続する。

【００３３】図１１に本実施の形態におけるＦＣ－ＡＬの接続形態を、ディスクドライブＤＫ１、ＤＫ２、ＤＫ３、ＤＫ４の接続形態を例として示す。各ディスクドライブは、それぞれＮＬポートを２個有する。各ディスクドライブの入出力ポートおよびスイッチＳＷ１、ＳＷ２の入出力ポートは、送信機Ｔｘと受信機Ｒｘを有する。スイッチＳＷ１、ＳＷ２のディスクアレイＤＡ１接続側入出力ポートは、ＦＬポートである。チャネルＤ１１により、スイッチＳＷ１、ディスクドライブＤＫ１、ＤＫ２、ＤＫ３、ＤＫ４をループ状に接続する。同様にチャネルＤ２１により、スイッチＳＷ２、ディスクドライブＤＫ１、ＤＫ２、ＤＫ３、ＤＫ４をループ状に接続する。これら２個のループは、ファイバチャネルのパブリックループであり、ディスクドライブＤＫ１、ＤＫ２、ＤＫ３、ＤＫ４は、スイッチＳＷ１またはＳＷ２を介してディスクアダプタＤＫＡ１またはＤＫＡ２と通信可能である。以上、チャネルＤ１１、Ｄ２１の接続形態を例に説明したが、チャネルＤ１２、Ｄ１３、Ｄ１４、Ｄ２２、Ｄ２３、Ｄ２４でも同様である。本実施の形態のディスク装置においてＲＡＩＤシステムを構築する場合は、ディスクグループＲ１、Ｒ２、Ｒ３、Ｒ４を、それぞれＲＡＩＤグループとする。本実施の形態では、４個のディスクドライブでＲＡＩＤグループを構成しているが、ＲＡＩＤグループを構成するドライブ数を４個に限るものではない。

【００３４】ディスクアレイＤＡ１内の全ディスクドライブは、ディスクアダプタＤＫＡ１およびＤＫＡ２のどちらからでもアクセス可能である。本実施の形態のディスク装置は、チャネルＤ１ｂ、Ｄ２ｂをスイッチＳＷ１、ＳＷ２故障時の迂回経路として使用する。例えばスイッチＳＷ１が故障した場合でも、ディスクアダプタＤＫＡ１はチャネルＤ１ｂとスイッチＳＷ２経由でディスクアレイＤＡ１にアクセスできる。逆に、スイッチＳＷ２が故障した場合は、ディスクアダプタＤＫＡ２はチャネルＤ２ｂとスイッチＳＷ１経由でディスクアレイＤＡ１にアクセスできるので、信頼性が高いディスク装置が実現できる。

【００３５】図１２に本発明の、第４の実施の形態であるディスク装置の構成を示す。本実施の形態のディスク装置は、第３の実施の形態のディスク装置に対して、スイッチＳＷ１、ＳＷ２間を接続するチャネルＤ３ａ、Ｄ３ｂを設けた点に特徴がある。本実施の形態において、ディスクアダプタＤＫＡ１、ＤＫＡ２と、スイッチＳＷ１、ＳＷ２と、ディスクアレイＤＡ１との間のデータ転送方式は、ファイバチャネルを使用している。本実施の形態のディスク装置は、ディスクコントローラＤＫＣ１、ＤＫＣ２と、スイッチＳＷ１、ＳＷ２と、ディスクアレイＤＡ１からなる。スイッチＳＷ１とＳＷ２は、第１の実施の形態と同様に速度変換と多重化、逆多重化を行うスイッチとして機能する。ディスクコントローラＤＫＣ１は、チャネルアダプタＣＨＡ１と、キャッシュメモリＣＭ１と、ディスクアダプタＤＫＡ１からなる。ディスクコントローラＤＫＣ２は、チャネルアダプタＣＨＡ２と、キャッシュメモリＣＭ２と、ディスクアダプタＤＫＡ２からなる。ディスクアダプタＤＫＡ１とスイッチＳＷ１をチャネルＤ１ａで接続し、ディスクアダプタＤＫＡ２とスイッチＳＷ２をチャネルＤ２ａで接続し、ディスクアダプタＤＫＡ１とスイッチＳＷ２をチャネルＤ１ｂで接続し、ディスクアダプタＤＫＡ２とスイッチＳＷ１をチャネルＤ２ｂで接続する。さらに、スイッチＳＷ１とＳＷ２をチャネルＤ３ａ、Ｄ３ｂで接続する。

【００３６】ディスクアレイＤＡ１を構成するディスクドライブは、入出力ポートを２個有する。例えば、ディスクドライブＤＫ１、ＤＫ２、ＤＫ３、ＤＫ４は、チャネルＤ１１およびＤ２１の両チャネルと接続する。ディスクアレイＤＡ１は、チャネルＤ１１とＤ２１に接続した４個のディスクからなるディスクアレイと、Ｄ１２とＤ２２に接続した４個のディスクからなるディスクアレイと、Ｄ１３とＤ２３に接続した４個のディスクからなるディスクアレイと、Ｄ１４とＤ２４に接続した４個のディスクからなるディスクアレイ、からなる。チャネルＤ１１、Ｄ１２、Ｄ１３、Ｄ１４、Ｄ２１、Ｄ２２、Ｄ２３、Ｄ２４は、図１１に示すようにＦＣ－ＡＬでディスクドライブを接続する。ディスクアレイＤＡ１内の全ディスクドライブは、ディスクアダプタＤＫＡ１およびＤＫＡ２のどちらからでもアクセス可能である。本実施の形態のディスク装置においてＲＡＩＤシステムを構築する場合は、ディスクグループＲ１、Ｒ２、Ｒ３、Ｒ４を、それぞれＲＡＩＤグループとする。本実施の形態では、４個のディスクドライブでＲＡＩＤグループを構成しているが、ＲＡＩＤグループを構成するドライブ数を４個に限るものではない。

【００３７】ディスクアダプタＤＫＡ１、ＤＫＡ２とディスクアレイＤＡ１のアクセス経路について、先ず、定常時（スイッチ故障無しの場合）について説明する。ディスクアダプタＤＫＡ１は、チャネルＤ１ａとＳＷ１を介してディスクアレイＤＡ１にアクセスする経路（経路１）と、チャネルＤ１ｂとスイッチＳＷ２とチャネルＤ３ａとスイッチＳＷ１を介してディスクアレイＤＡ１にアクセスする経路（経路２）を有する。同様に、ディスクアダプタＤＫＡ２は、チャネルＤ２ａとＳＷ２を介してディスクアレイＤＡ１にアクセスする経路（経路３）と、チャネルＤ２ｂとスイッチＳＷ１とチャネルＤ３ｂとスイッチＳＷ２を介してディスクアレイＤＡ１にアクセスする経路（経路４）を有する。一方、スイッチ故障時は、チャネルＤ１ｂ、Ｄ２ｂを迂回経路として使用する。例えばスイッチＳＷ１が故障した場合でも、ディスクアダプタＤＫＡ１はチャネルＤ１ｂとスイッチＳＷ２経由でディスクアレイＤＡ１にアクセスできる。逆に、スイッチＳＷ２が故障した場合は、ディスクアダプタＤＫＡ２はチャネルＤ２ｂとスイッチＳＷ１経由でディスクアレイＤＡ１にアクセスできる。

【００３８】次に、本実施の形態におけるディスクアダプターディスクアレイ間のスループットについて説明する。例として、チャネルＤ１ａ、Ｄ１ｂ、Ｄ２ａ、Ｄ２ｂ、Ｄ３ａ、Ｄ３ｂ上のデータ伝送速度をチャネル当り２Ｇｂｐｓとし、チャネルＤ１１、Ｄ１２、Ｄ１３、Ｄ１４、Ｄ２１、Ｄ２２、Ｄ２３、Ｄ２４上のデータ伝送速度をチャネル当り１Ｇｂｐｓであるとする。このとき、スイッチＳＷ１とディスクアレイＤＡ１間の総スループットは４Ｇｂｐｓである。ディスクアダプタＤＫＡ１とスイッチＳＷ１間は、上記経路１および経路２でアクセスすることにより、総スループットは４Ｇｂｐｓとなる。スイッチＳＷ１のディスクアダプタＤＫＡ１側と、ディスクアレイＤＡ１側のスループットが共に４Ｇｂｐｓであるので、ディスクアダプタＤＫＡ１とディスクアレイＤＡ１間のスループットは４Ｇｂｐｓとなる。同様に、スイッチＳＷ２とディスクアレイＤＡ１間の総スループットは４Ｇｂｐｓである。ディスクアダプタＤＫＡ２とスイッチＳＷ２間は、上記経路３および経路４でアクセスすることにより、総スループットは４Ｇｂｐｓとなる。スイッチＳＷ２のディスクアダプタＤＫＡ２側と、ディスクアレイＤＡ１側のスループットが共に４Ｇｂｐｓであるので、ディスクアダプタＤＫＡ２とディスクアレイＤＡ２間のスループットは４Ｇｂｐｓとなる第３の実施の形態（図８）において、上記のチャネル当りスループット値を適用すると、チャネルＤ１ｂ、Ｄ２ｂをスイッチ故障時の迂回経路としか使用しないので、ディスクアダプタＤＫＡ１とディスクアレイＤＡ１間のスループットは、チャネルＤ１ａ上のスループットに制限され、２Ｇｂｐｓとなる。同様に、ディスクアダプタＤＫＡ２とディスクアレイＤＡ１間のスループットは、チャネルＤ２ａ上のスループットに制限され、２Ｇｂｐ

ｓとなる。第３の実施の形態において、ディスクアダプターディスクアレイ間スループットを４Ｇｂｐｓにするためには、チャネルＤ１ａおよびＤ２ａのデータ伝送速度を、それぞれ４Ｇｂｐｓに高める必要がある。以上から、本実施の形態によれば、ディスクアダプタースイッチ間のチャネル当りデータ伝送速度が低くても、ディスクアダプターディスクアレイ間の総スループットが高いディスク装置が実現できる。

【００３９】

【発明の効果】以上説明したように、本発明によれば以下の効果がある。ディスクアダプタとディスクアレイ間のスループットが高いディスク装置を提供できる。また、ディスクアダプタとディスクアレイ間のスループットが高く、且つディスクドライブ接続台数が多いディスク装置を提供できる。また、信頼性の高いディスクアレイを有するディスク装置を提供できる。また、信頼性が高いディスクアダプタとディスクアレイ間ネットワークを有するディスク装置を提供できる。また、信頼性およびスループットが高いディスクアダプタとディスクアレイ間ネットワーク、を有するディスク装置を提供できる。また、ディスクからの読み出しおよびディスクへの書きこみを高スループット化できるディスク装置を提供できる。また、高スループットを維持できるディスク装置を提供できる。また、ディスクアダプタとディスクアレイ間のスループットが高く低コストなディスク装置を提供できる。

【図面の簡単な説明】

【図１】本発明の第１の実施の形態のディスク装置を示す図である。

【図２】本発明に用いるスイッチの構成を示す図である。

【図３】本発明に用いるスイッチの構成を示す図である。

【図４】本発明に用いるスイッチの動作を示す図である。

【図５】本発明に用いるスイッチの動作を示す図である。

【図６】本発明第１の実施の形態に対して、ディスクドライブを増設する方法を示す図である。

【図７】本発明の第２の実施の形態のディスク装置を示す図である。

【図８】本発明の第３の実施の形態のディスク装置を示す図である。

【図９】従来のディスク装置を示す図である。

【図１０】ＦＣ－ＡＬによる接続形態を説明する図である。

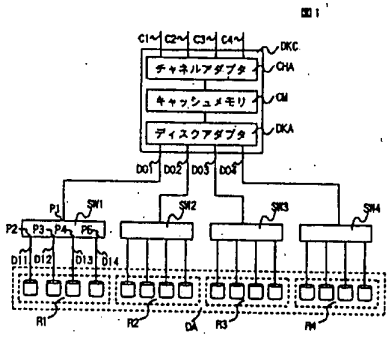
【図１１】ＦＣ－ＡＬによる接続形態を説明する図である。

【図１２】本発明の第４の実施の形態のディスク装置を示す図である。

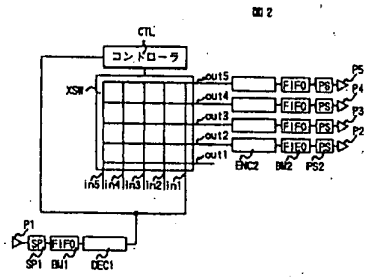
【符号の説明】

ＤＫＣ、ＤＫＣ１、ＤＫＣ２　ディスクコントローラ
ＣＨＡ、ＣＨＡ１、ＣＨＡ２　チャネルアダプタ
ＣＭ、ＣＭ１、ＣＭ２　キャッシュメモリ
ＤＫＡ、ＤＫＡ１、ＤＫＡ２　ディスクアダプタ
ＤＡ、ＤＡ１～ＤＡ４　ディスクアレイ
ＤＫ１～ＤＫ４　ディスクドライブ
Ｒ１～Ｒ６　ディスクグループ
Ｃ１～Ｃ４、Ｄ１～Ｄ４、Ｄ０１～Ｄ０５、Ｄ１１～Ｄ１４、Ｄ２１～Ｄ２４、Ｄ１ａ、Ｄ１ｂ、Ｄ２ａ、Ｄ２ｂ、Ｄ３ａ、Ｄ３ｂ　チャネル
ＳＷ１～ＳＷ４　スイッチ
Ｐ１～Ｐ５　スイッチポート
ＸＳＷ　クロスバスイッチ
ＣＴＬ　スイッチコントローラ
ｉｎ１～ｉｎ５　クロスバスイッチ入力ポート
ｏｕｔ１～ｏｕｔ５　クロスバスイッチ出力ポート
ＳＰ１、ＳＰ２　シリアルパラレル変換装置
ＰＳ１、ＰＳ２　パラレルシリアル変換装置
ＢＭ１、ＢＭ２　バッファメモリ
ＤＥＣ１、ＤＥＣ２　８Ｂ１０Ｂ変換デコーダ
ＥＮＣ１、ＥＮＣ２　８Ｂ１０Ｂ変換エンコーダ
Ｔ１、Ｔ２、Ｔ３、
Ｔ４、Ｔ５、Ｔ６　フレームの時間
Ｔｘ　送信機
Ｒｘ　受信機
ＮＬ　ＮＬポート
ＦＬ　ＦＬポート

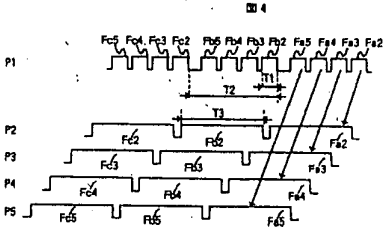
【図１】



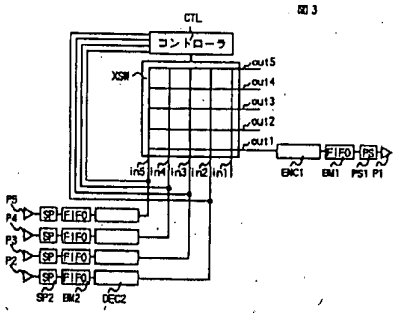
【図２】



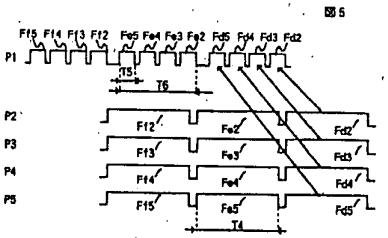
【図４】



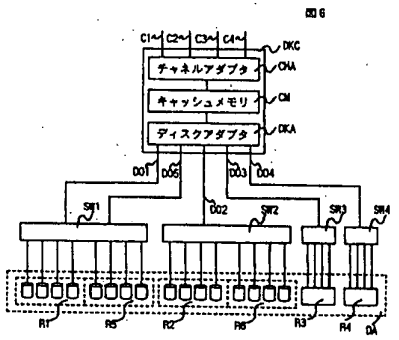
【図３】



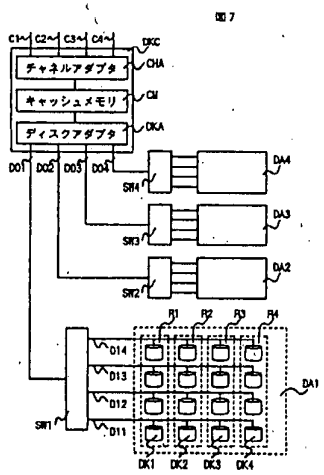
【図５】



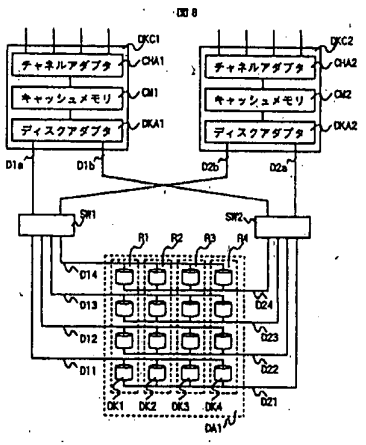
【図６】



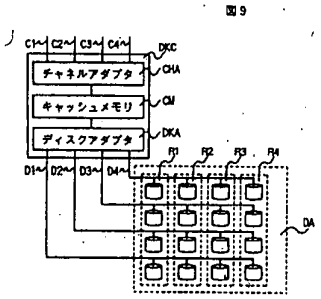
【図７】



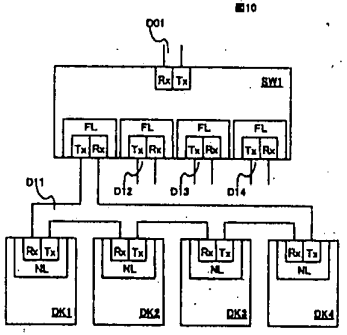
【図８】



【図９】

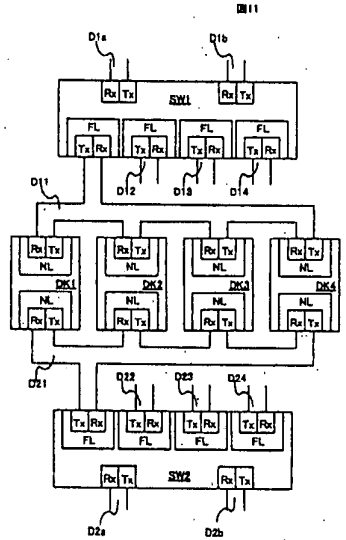


【図１０】

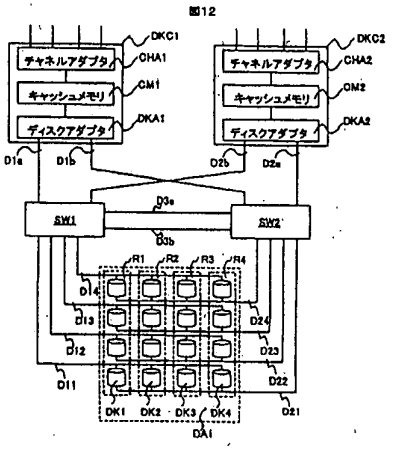


ディスクアダプタとディスクアレイをスイッチを介して接続したディスク装置　特開２００３－３０３０５５

【図１１】



【図１２】



Simplify™

Network Storage Group
Host Products
Technology Brief
October 8, 2002

FULL-DUPLEX AND FIBRE CHANNEL

WHAT IS FULL-DUPLEX?

With full-duplex data transmission, data is received and transmitted at the same time. A Fibre Channel adapter that has full-duplex capability can send data to a Fibre Channel node and receive data from that node simultaneously.

WHY FIBRE CHANNEL?

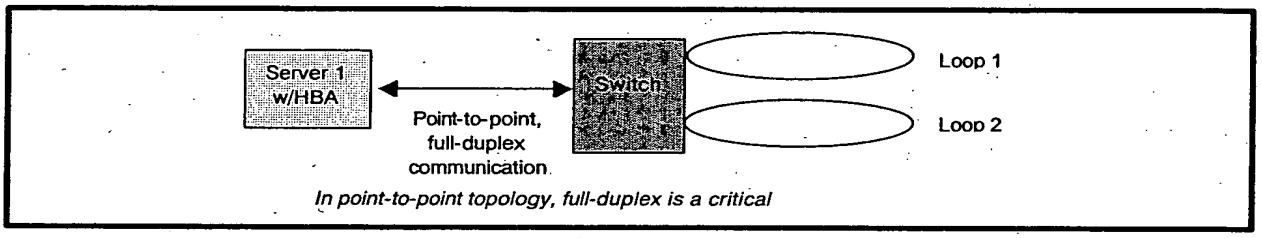
Traditional I/O channels are buses. Buses are like one-way tunnels. Parallel SCSI and ATA, for example, can only process a single point-to-point transfer at a time. Fibre Channel is different. On a single Fibre Channel cable, there are two connections between any two devices:

- The outbound half of the cable goes from the transmitting device to the receiving device.
- The inbound half of the cable goes from the receiver back to the sender and completes the connection.

These connections are physically separate; Fibre Channel supports separate communications being in process on each half of the cable at the same time. This capability, called full-duplex communication, makes Fibre Channel more efficient than traditional buses.

WHEN IS FULL-DUPLEX MOST CRITICAL?

Full-duplex is most important in point-to-point communications on a switched fabric. A host system communicating with a switch can take advantage of the simultaneous send and receive capabilities. A switch is most likely to have data ready to transmit to the host when a connection is opened between them.



In point-to-point topology, full-duplex is a critical

In the above diagram, the host adapter in the Server 1 is communicating simultaneously with a drive on Loop 1 and a drive on Loop 2. (The drives can be spread across more than two loops; however, only two loops are needed to illustrate the feature.) In this configuration, the system can realize the performance potential of Fibre Channel full duplex capability.

SUMMARY

Full-duplex is one of many features that highlight the advantages of Fibre Channel. When a host adapter accesses drives on multiple loops on a switch and has a workload that keeps the attached drives busy, using full-duplex results in a considerable improvement in performance.